UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA
E INVESTIGAÇÃO OPERACIONAL

Estimação em Pequenos Domínios no âmbito do Inquérito ao Emprego

Soraia Alexandra Gonçalves Pereira

MESTRADO EM ESTATÍSTICA

2012

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA
E INVESTIGAÇÃO OPERACIONAL

Estimação em Pequenos Domínios no âmbito do Inquérito ao Emprego

Soraia Alexandra Gonçalves Pereira

Dissertação orientada pela Prof. Doutora Maria Lucília Carvalho
e co-orientada pelo Dr. Luís Correia

MESTRADO EM ESTATÍSTICA

2012

Agradecimentos

Em primeiro lugar, gostaria de agradecer aos meus orientadores: Prof. Dra. Lucília Carvalho e Dr. Luís Correia. Sem a ajuda de ambos, não seria possível fazer este trabalho. Agradeço à Prof. Lucília essencialmente por toda a dedicação, conselhos e sugestões, e pela disponibilidade para a discussão de questões que, para mim, não eram tão simples. A sua experiência e perspicácia ajudaram em muito este trabalho. Agradeço ao Dr. Luís Correia, por toda a motivação que me transmitiu para a área dos pequenos domínios e por tudo o que me ensinou. Foi e continua a ser uma forte orientação. A sua capacidade para a resolução de problemas práticos e a experiência na área, permitiram que este trabalho fosse possível.

Devo agradecer também às instituições INE e FCT. Ao INE por me acolher e motivar para este problema, e à FCT por fazer com que a ligação ao INE fosse possível, através de uma bolsa de investigação.

Agradeço à Prof. Helena Iglésias pela disponibilidade para responder a questões relativas à escrita de uma dissertação, e agradeço à Prof. Antónia Turkman por tudo o que me ensinou com o software R, pois essa aprendizagem foi fulcral para a produção dos resultados.

Gostaria de agradecer também à Paula Marques e à Mafalda Cabral, técnicas superiores do Departamento de Metodologia e Sistemas de Informação do INE, por fornecerem prontamente os ficheiros de dados necessários ao estudo feito, e alguns apuramentos.

Agradeço ao Pedro Campos, técnico superior do Departamento de Metodologia e Sistemas de Informação do INE na Delegação do Porto, por toda a ajuda e interesse demonstrado.

Agradeço ao Davide Miranda, bolseiro de investigação do Departamento das Estatísticas Sociais do INE, pela ajuda com os mapas apresentados na dissertação.

À Sofia e à Soraia, bolseiras de investigação do Departamento das Contas Nacionais do INE, agradeço a ajuda com as traduções, e a força e motivação dadas na última fase do trabalho.

Ao Luciano, um agradecimento especial por todo o carinho, compreensão, ajuda, e motivação.

Um obrigado a todos os meus amigos, em especial à Marta, Rita, e Margarida, pela vossa

amizade constante e apoio dado durante todo o processo.
Um agradecimento muito especial à minha família, por estar sempre comigo.

Abstract

Measurement of unemployment takes an enormous social and political importance in contemporary societies. In Portugal, the INE (Instituto Nacional de Estatística), as the entity responsible for production and dissemination of national statistics, produces national estimates for the unemployment rate, from a direct estimation method. These estimates are released by geographic levels known as NUTS I and NUTS II. According to the European nomenclature, territorial levels used for statistical purposes are designated by NUTS I, II or III, corresponding to a growing disaggregation order. Currently, there is an increasingly need of knowing unemployment values for more disaggregated levels. However and as expected, when working with a sample, in this case, the Labour Force Survey sample, the higher the level of disaggregation required, the greater the variance of the estimates that are calculated by the direct method. INE believes that estimates produced by this method for the NUTS III do not achieve enough precision to be published. To overcome this problem the present study develops a methodology, to construct with an acceptable accuracy, quarterly and annual unemployment rate estimates for this geographical level, and that, when aggregated, agree with the official estimates for NUTS II. These estimates were produced from a regression model-based on a methodology used by the Office for National Statistics (ONS) from UK, and using auxiliary information from the Employment and Professional Training Institute (IEFP). Quarterly estimates are produced for the 1st quarter of 2011 and afterwards annual estimates are produced for 2011. The corresponding coefficients of variation are also calculated using the Bootstrap methodology.

Keywords: Employment Survey, Unemployment rate, Small Area Estimation, Logistic Regression, Office for National Statistics, Bootstrap.

Resumo

A quantificação do desemprego assume uma enorme importância social e política nas sociedades contemporâneas. Em Portugal, o Instituto Nacional de Estatística (INE), enquanto entidade responsável pela produção e divulgação das estatísticas nacionais, produz estimativas para a taxa de desemprego para todo o país, a partir de um método de estimação direto. Estas estimativas são divulgadas por níveis geográficos designados por NUTS I e NUTS II. Os níveis territoriais usados para fins estatísticos tomam, de acordo com a nomenclatura europeia, o nome de NUTS, ao qual se junta a numeração I, II ou III segundo uma ordem de desagregação crescente. Atualmente há uma necessidade de se conhecer valores do desemprego a níveis cada vez mais desagregados. Contudo, como é de esperar, com base numa amostra, neste caso na amostra do Inquérito ao Emprego, quanto maior o nível de desagregação pretendido, maior é a variância das estimativas calculadas pelo método direto. O INE entende que as estimativas produzidas para as NUTS III, usando este método, não têm precisão suficiente para serem publicadas. Para ultrapassar este problema, este estudo desenvolve uma metodologia de construção de estimativas trimestrais e anuais da taxa de desemprego para este nível geográfico, com uma precisão considerada aceitável, e que, quando agregadas, coincidem com as estimativas oficiais para as NUTS II. Estas estimativas foram produzidas a partir de um modelo de regressão, com base numa metodologia utilizada pelo *Office for National Statistics* (ONS), do Reino Unido, utilizando informação auxiliar proveniente do Instituto do Emprego e Formação Profissional (IEFP). São produzidas estimativas trimestrais para o 1º trimestre de 2011 e depois estimativas anuais para 2011. Os correspondentes coeficientes de variação são também calculados, usando a metodologia Bootstrap.

Palavras chave: Inquérito ao Emprego, Taxa de desemprego, Estimação em Pequenos Domínios, Regressão Logística, *Office for National Statistics*, Bootstrap.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O Inquérito ao Emprego (IE) em Portugal é uma operação estatística conduzida pelo INE, que incide sobre uma amostra da população. O objetivo principal deste inquérito é obter estimativas trimestrais da população ativa (empregados e desempregados) e inativa. Com base nestas estimativas é possível calcular um vasto número de indicadores, entre eles a taxa de emprego e a taxa de desemprego. Estes dois indicadores são muito utilizados para caracterizar a população face ao mercado de trabalho. Enquanto que a taxa de emprego mede a proporção de pessoas ativas que estão empregadas, a taxa de desemprego, que é o indicador de interesse neste estudo, mede a proporção de pessoas ativas que estão desempregadas.

As estimativas da taxa de desemprego são publicadas trimestralmente pelo INE ao nível regional NUTS II (Norte, Centro, Lisboa, Alentejo, Algarve, Região Autónoma dos Açores, Região Autónoma da Madeira). Estas características são calculadas usando um método direto e estão disponíveis por região, sexo, idade, estrutura familiar, e seus cruzamentos.

Atualmente, as necessidades de conhecimento do mercado de trabalho impõem estimativas fiáveis para o total de desempregados e para a taxa de desemprego a níveis mais desagregados, nomeadamente ao nível NUTS III.

As NUTS III são domínios de menor dimensão em que se subdividem as NUTS II e, como tal, a informação sobre algumas das variáveis de interesse não é suficiente para se obter estimativas com precisão aceitável recorrendo ao método já referido.

O objetivo deste estudo é encontrar um método que permita obter estimativas trimestrais e anuais dos totais de desemprego e da taxa de desemprego ao nível NUTS III, com uma boa precisão. Não existe uma regra específica que permita definir se determinada estimativa não tem uma boa precisão ou se esta é aceitável. A decisão, tomada por uma instituição, de divulgar ou não estimativas com determinada precisão tem, na maior parte

das vezes, um caráter subjetivo. No entanto, é quase consensual que estimativas com coeficientes de variação superiores a 20% não devem ser publicadas.

Este problema enquadra-se na *Estimação em Pequenos Domínios*. O termo pequeno domínio é usado para denotar qualquer subpopulação para a qual não é possível produzir estimativas diretas com precisão adequada. Na construção de estimativas para pequenos domínios com um bom nível de precisão, é necessário usar estimadores indiretos. Estes estimadores usam "força emprestada" por outros domínios adjacentes através de valores da variável de interesse, e assim pode considerar-se que o tamanho "efetivo" da amostra aumenta. Ou seja, estes estimadores usam informação fora do domínio de interesse para o qual se pretende produzir as estimativas.

A existência de boa informação auxiliar e a determinação de modelos adequados são ferramentas cruciais para a formação de estimadores indiretos. No caso que estamos a tratar, temos como informação auxiliar os dados do IEFP (Instituto do Emprego e Formação Profissional).

Mensalmente são divulgados dados do desemprego registado nos Centros de Emprego (CE). A contagem dos desempregados inscritos nos CE é baseada nos registos administrativos dos indivíduos que se inscrevem nos referidos Centros, com o objetivo de obter um emprego por conta de outrem. Estes indivíduos estão desempregados e imediatamente disponíveis para trabalhar.

Chama-se à atenção que o conceito de desempregado utilizado no IE, que será definido no capítulo *Inquérito ao Emprego* é diferente do conceito de desempregado do IEFP e, portanto, os números não têm que coincidir. Os indivíduos desempregados para o IE podem não estar registados nos CE. Também pode acontecer que indivíduos inscritos nos CE não sejam considerados desempregados pelo IE se, por exemplo, estes não procuraram trabalho no período de referência ou nas três semanas anteriores. No entanto, é evidente que existe uma interseção entre os conjuntos de indivíduos a que se referem estes dois conceitos.

O INE tem acesso a esta informação proveniente do IEFP por concelho, sexo, e pelos grupos etários 15-24, 25-34, 35-49, 50+. A utilização desta informação é preciosa para a estimação para as NUTS III, especialmente porque torna possível a utilização de modelos de regressão.

De facto, os modelos mais utilizados para resolver o problema de estimação em pequenos domínios são baseados na regressão linear. Entre estes modelos, destaca-se o modelo usado pelo *Office for National Statistics* (ONS), Instituto de Estatística do Reino Unido, que aplicou este modelo para produzir estimativas do desemprego ao nível de domínios territoriais designados por *Parliamentary Constituencies* (PCs). Este será o modelo que utilizaremos neste trabalho.

Este trabalho divide-se em 6 capítulos.
No segundo capítulo é feita uma abordagem à *estimação em pequenos domínios* e referência a alguns dos métodos já desenvolvidos para tratar este problema.
No capítulo *Inquérito ao Emprego* é descrita a metodologia de amostragem, estimação, e precisão, que está por detrás deste inquérito. São definidos os conceitos de desempregado, população ativa, e taxa de desemprego, utilizados no inquérito.
O capítulo seguinte descreve a metodologia usada pelo ONS e a sua aplicação aos dados do IE. Antes da aplicação do modelo escolhido, é feita a sua contextualização aos modelos de regressão. Dar-se-á destaque aos modelos multinível mistos de regressão logística, uma vez que é neste tipo de modelos que o modelo do ONS se insere.
No capítulo *Análise dos Resultados* são apresentadas as estimativas obtidas e calculados os indicadores de performance. O modelo é aplicado aos dados do primeiro trimestre de 2011 do Inquérito ao Emprego. Para além das estimativas trimestrais, também são feitas estimativas anuais. Será constituída a amostra anual de 2011 (desde o primeiro trimestre de 2011 até ao quarto trimestre) através de um esquema de rotação, e são feitas as estimativas com base no mesmo modelo utilizado para produzir estimativas trimestrais. A informação auxiliar utilizada corresponde à média dos dados do IEFP de doze meses correspondentes ao ano em causa.
Como é evidente, as estimativas têm um erro associado que deve ser conhecido, mas, para além de uma boa precisão, é ainda exigido que as estimativas sejam consistentes, isto é, que a soma das estimativas dos totais de desemprego das NUTS III que pertencem a determinada região NUTS II seja igual à estimativa publicada para essa região. Iremos ver que a consistência das estimativas pode ser assegurada através de uma calibração.
Por fim, são feitas as estimativas da taxa de desemprego ao nível NUTS III. Os coeficientes de variação das estimativas são estimados pelo método de Bootstrap.
No último capítulo são apresentadas as conclusões deste estudo.

Capítulo 2

Estado da Arte das Metodologias de Estimação em Pequenos Domínios

2.1 Trabalho desenvolvido

A procura de estimativas fiáveis em domínios específicos para indicadores com as mais variadas utilizações é uma necessidade em franco crescimento, particularmente em contextos de exclusão social, bem-estar social, e recessão económica.
Estes domínios podem ser definidos como áreas geográficas, subgrupos sócio-económicos ou outras subpopulações, consoante o contexto.
Para satisfazer esta procura, os responsáveis pela produção de estatísticas tentam construir estimadores baseados nas amostras disponíveis com a melhor precisão possível. No entanto, em alguns domínios, a dimensão da amostra correspondente é tão pequena, que não permite obter estimativas fiáveis através de métodos diretos. Estes domínios são conhecidos como *pequenos domínios* [1].
A estimativa num domínio é *direta* se se baseia apenas em informação específica da amostra que está nele contida. Pode usar informação de uma ou mais variáveis auxiliares relacionadas com a variável de interesse, mas sempre dentro de um domínio.
Quando se tratam *pequenos domínios*, é necessário recorrer a estimadores *indiretos*. Estes usam *informação emprestada* por outros domínios, como forma de aumentar a informação contida na amostra do domínio em causa.
A estimação *design-based* considera que os valores tomados pela variável de interesse, bem como os parâmetros que lhe estão associados, são constantes e que a componente probabilística é introduzida pelo plano de amostragem.
A estimação *model-based* considera que os valores tomados pela variável de interesse resultam da realização de variáveis aleatórias, cuja distribuição conjunta é especificada por

um modelo, sendo que os parâmetros associados têm distribuições dela derivadas. Para estimativas em pequenos domínios, concluiu-se que, em geral, os estimadores model-based produzem melhores resultados.

As inferências de estimadores model-based são feitas com base na distribuição definida pelo modelo assumido e, como tal, a seleção e validação do modelo é um processo crucial para este tipo de estimação. Se os modelos assumidos não se ajustarem bem aos dados, os estimadores serão enviesados, o que pode levar a inferências erradas.

Para variáveis de interesse que são binárias ou contagens, os modelos lineares generalizados mistos são os mais adequados. Em particular, no caso binário, usam-se modelos logísticos lineares mistos, e no caso de contagens usam-se modelos loglineares mistos.

Os estimadores indiretos tradicionais são, na grande maioria, design-based e as suas variâncias (induzidas pelo desenho da amostra) são usualmente pequenas relativamente às variâncias dos estimadores diretos. No entanto, estes estimadores são muito frequentemente enviesados, e este enviesamento não decresce com o aumento da amostra.

Os modelos baseados em efeitos aleatórios específicos da área, que têm em consideração a variação entre as áreas, e que são explicados por variáveis auxiliares ou covariáveis, são chamados *small area models* [1].

Os modelos para pequenas áreas podem ser classificados em dois tipos: 1) *Modelos ao nível de área* (ou agregado) que relacionam a variável de interesse com as covariáveis, por área. 2) *Modelos ao nível de unidade* que relacionam a variável de interesse com as covariáveis, por unidade. Neste trabalho, irá ser apresentado um modelo ao nível de área, visto que não há informação auxiliar disponível ao nível do indivíduo.

Têm sido propostos vários estimadores indiretos. No entanto, há ainda algumas reservas quanto ao comportamento destes estimadores quando aplicados a situações reais. Para tentar resolver este problema, foram criados vários projetos, entre eles o projeto EURAREA, que teve como objetivo fazer um estudo que permitisse ao Eurostat e aos INEs Europeus decidir se deviam aplicar estes estimadores, e qual o modo de o fazer. Este projeto foi financiado pelo Eurostat no âmbito do Quinto Programa da União Europeia e decorreu entre Janeiro de 2001 e Junho de 2004, tendo terminado em Fevereiro de 2005.

Foram testados, por simulação, mais de 20 estimadores, sendo que 7 deles mereceram especial atenção: o estimador direto, o estimador de regressão generalizado, o estimador sintético a nível micro (da unidade), o estimador sintético a nível macro (da área), o estimador sintético sob o modelo de regressão logística, o estimador EBLUP a nível micro, e o estimador EBLUP a nível macro. A programação foi desenvolvida no software SAS (Statistical Analysis System).

Foi também criado um instrumento essencial, o Cenex. Foi proposto em Setembro de 2002 com o objetivo de reunir conhecimentos neste domínio de todas as organizações do

Sistema Estatístico Europeu (SEE) que, em inglês, corresponde a *European Statistical System* (ESS), beneficiando assim todo o sistema.
Em 2005, surgiu o SDC, um dos projetos Cenex, com o intuito de estimular o desenvolvimento de metodologias inovadoras e torná-las de fácil acesso a todos os Estados-Membros através do Eurostat.
Posteriormente, o nome Cenex foi substituído por ESSnet (Collaborative Networks ESS). Em Dezembro de 2009 começou um outro projeto importante do ESSnet, o SAE (Small Area Estimation), tendo terminado em Março de 2012. Este projeto teve como objetivo desenvolver um quadro que permitisse a produção de estimativas para áreas pequenas, no que diz respeito a inquéritos do ESS. Para o fazer, foram usadas várias linguagens de programação, entre elas SAS e R.
Um outro programa que também tenciona desenvolver procedimentos para estimativas em pequenos domínios é o Sample. No entanto, tem como principal objetivo identificar e desenvolver indicadores e modelos para a desigualdade e pobreza. É financiado pela Comissão Europeia e começou em 2008, com duração de três anos.
Têm sido feitos outros estudos para melhorar estimativas já produzidas para pequenos domínios. É o caso de um estudo recente feito pelo ISTAT (Instituto de Estatística da Itália) acerca do uso de informação espacial nos modelos de pequenas áreas para a estimação da taxa de desemprego ao nível sub-regional. Este estudo compara os modelos do projeto EURAREA com modelos logit. Em cada um destes modelos é incorporada informação espacial. O estudo de simulação permite concluir que o modelo logit com efeitos aleatórios e com a incorporação de informação espacial é o modelo que apresenta melhores resultados no que diz respeito à centricidade e variância. Tal como o ISTAT, o ONS também já produz estimativas ao nível sub-regional. Este Instituto usa um modelo logit com efeitos aleatórios para produzir estas estimativas. Mais à frente, iremos ver a descrição deste modelo.
Como se pode constatar, há um grande esforço por parte dos Institutos Nacionais de Estatística e do Eurostat para desenvolver software que possa ser utilizado na estimação de pequenos domínios com a melhor precisão possível.

2.2 Estimadores standard para totais

Considere-se uma população finita constituída por N elementos, que se denota por $U = \{1, ..., k, ..., N\}$. Seja y a variável de interesse, e y_k o valor de y para o k-ésimo elemento da população.

O total de y, t_y, é então dado por:

$$t_y = \sum_U y_k. \tag{2.1}$$

Pretende-se fornecer estimativas não só para toda a população U, mas também para subpopulações específicas, os domínios.
Seja U_d um domínio de interesse, com $U_d \subseteq U$, e N_d o número de elementos de U_d.
Considere-se a nova variável:

$$y_{dk} = \begin{cases} y_k & \text{se } k \in \mathrm{U}_d \\ 0 & \text{caso contrário} \end{cases}$$

Então t_{yd}, o total da variável de interesse no domínio U_d, pode ser escrito como

$$t_{yd} = \sum_{U_d} y_k = \sum_U y_{dk}.$$

De seguida são apresentados vários estimadores para este total.

1º - Estimador Direto

O estimador direto ou de Horvitz-Thompson [2] para toda a população U é dado por:

$$\hat{t}_y^{HT} = \sum_{k \epsilon s} w_k y_k,$$

em que s é uma amostra neste universo, $s \subseteq U$, e w_k é o inverso da probabilidade de inclusão do indivíduo k na amostra s. O valor w_k é chamado de "peso do indivíduo", visto que corresponde ao número de elementos que o indivíduo representa na população.
O estimador direto para o total no domínio U_d é dado por:

$$\hat{t}_{yd} = \sum_{k \epsilon s_d} w_k y_k,$$

em que s_d é a restrição da amostra ao domínio U_d.
Como este estimador não recorre a informação externa ao domínio de interesse, inclui-se nos estimadores diretos. É também um estimador design-based.

2º - Estimador Sintético

Suponhamos que temos q variáveis auxiliares. Denotemos o vetor de variáveis auxiliares ou covariáveis por $\boldsymbol{x} = (1, x_1, ..., x_j, ..., x_q)$. O valor da j-ésima variável para o k-ésimo

elemento da população é denotado por x_{jk}. Para o k-ésimo elemento, o correspondente vetor de covariáveis designa-se por $\boldsymbol{x_k} = (1, x_{1k}, ..., x_{jk}, ..., x_{qk})$.
Supõe-se que se observa $(y_k\ , \boldsymbol{x_k})$ para todos os elementos $k \in s$, mas que também se conhece $\boldsymbol{x_k}$ para $k \in (U - s)$.
Os estimadores sintéticos apoiam-se num modelo de regressão linear simples ao nível das unidades,

$$y_k = \boldsymbol{x_k}\boldsymbol{\beta} + \varepsilon_k,$$

onde $\beta = (\beta_0, \beta_1, ..., \beta_q)'$ é o vetor dos coeficientes de regressão.
Este modelo linear pode ser estocástico ou determinístico. É estocástico quando assumimos que y tem uma distribuição aleatória inerente, ou seja, estamos no âmbito da inferência model-based. Neste caso, o modelo ficará melhor descrito por

$$Y_k = \boldsymbol{x_k}\boldsymbol{\beta} + \varepsilon_k \tag{2.2}$$

em que ϵ_k são variáveis aleatórias com $E[\varepsilon_k] = 0$ e $var[\varepsilon_k] = \sigma$ (constante) ou $var[\varepsilon_k] = \sigma_k$. Então, tem-se

$$E[Y_k] = \boldsymbol{x_k}\boldsymbol{\beta}$$

e β pode ser estimado pelo método dos mínimos quadrados generalizados (quer a variância dos erros seja constante ou não).
O estimador dos mínimos quadrados generalizados é dado por:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{B} = \left(\sum_{k\epsilon U} \frac{\boldsymbol{x_k}\boldsymbol{x_k'}}{\sigma_k^2}\right)^{-1} \left(\sum_{k\epsilon U} \frac{\boldsymbol{x_k} y_k}{\sigma_k^2}\right).$$

Uma vez que não se conhecem os valores de y_k para todo o universo, $\boldsymbol{B}$ terá que ser estimado à custa da amostra e teremos uma segunda estimativa (agora proveniente do desenho da amostra)

$$\hat{\boldsymbol{B}} = \left(\sum_{k\epsilon s} a_k \boldsymbol{x_k}\boldsymbol{x_k'}\right)^{-1} \left(\sum_{k\epsilon s} a_k \boldsymbol{x_k} y_k\right).$$

Estamos a supor que o modelo é válido para todo o universo, logo a estimativa sintética do total num domínio U_d será dada por:

$$\hat{t}_{yd}^{syn} = \boldsymbol{t_{xd}}\hat{\boldsymbol{B}}$$

em que

$$\boldsymbol{t_{xd}} = (t_{x_1 d}, t_{x_2 d}, ..., t_{x_q d})$$

é o total dentro do domínio para cada covariável $x_1, ..., x_q$.
Desta forma, o estimador é indireto, pois o estimador de $\boldsymbol{B}$ baseia-se em s contendo assim informação de outros domínios além de U_d.

3º - Estimadores Compostos

Para tentar melhorar as propriedades de dois estimadores constroem-se muitas vezes estimadores compostos que são uma combinação linear convexa dos dois estimadores.
Isso acontece com o estimador sintético (enviesado, mas de pequena variância), que é composto frequentemente com o estimador direto (centrado, mas de grande variância),

$$\hat{t}_{yd}^{com} = \gamma_d \hat{t}_{yd} + (1-\gamma_d)\hat{t}_{yd}^{syn},\ 0 < \gamma_d < 1.$$

4º - Estimador de Regressão Generalizado (GREG)

A construção do estimador de regressão generalizado faz uso da estimação direta e do modelo de regressão linear usado nos estimadores sintéticos. Começa por se basear na seguinte igualdade:

$$y_k = y_k^0 + (y_k - y_k^0)$$

em que y_k^0 é uma aproximação qualquer de y_k, conhecida para todos os $k \epsilon U$.
O total num domínio U_d pode então ser escrito como:

$$t_{yd} = \sum_{U_d} y_k = \sum_{U_d} y_k^0 + \sum_{U_d} (y_k - y_k^0).$$

Note-se que os y_k só são conhecidos para s, ao contrário do que acontece com a aproximação y_k^0 que é conhecida para todos os elementos de U (logo para todo o U_d).
Recorrendo ao estimador direto para estimar estas diferenças, obtem-se

$$\hat{t}_{yd} \approx \sum_{U_d} y_k^0 + \sum_{s_d} w_k (y_k - y_k^0).$$

Supondo que a aproximação para cada y_k^0 é feita através do modelo linear 2.2, temos

$$\hat{t}_{yd}^{GREG} = \sum_{U_d} \boldsymbol{x_k}\hat{\boldsymbol{B}} + \sum_{s_d} a_k (y_k - \boldsymbol{x_k}\hat{\boldsymbol{B}}).$$

Fazendo as contas, vem

$$\hat{t}_{yd}^{GREG} = \boldsymbol{t_{xd}}\hat{\boldsymbol{B}} + \hat{t}_{yd} - \hat{\boldsymbol{t}}_{\boldsymbol{xd}}\hat{\boldsymbol{B}}$$

em que $\hat{\boldsymbol{t}}_{\boldsymbol{xd}} = (\hat{t}_{x_1 d}, ..., \hat{t}_{x_q d})$ é a estimativa direta do total no domínio U_d para cada uma das covariáveis. Trocando a ordem das parcelas, tem-se

$$\hat{t}_{yd}^{GREG} = \hat{t}_{yd} + (\boldsymbol{t_{xd}} - \hat{\boldsymbol{t}}_{\boldsymbol{xd}})\hat{\boldsymbol{B}}.$$

Note-se que este estimador também pode ser interpretado como o estimador direto adicionado de uma correção que será tanto mais pequena quanto a estimação direta for mais adequada $\boldsymbol{t_{xd}} \approx \hat{\boldsymbol{t}}_{\boldsymbol{xd}}$.
Tanto o estimador sintético como o GREG são enviesados mas têm normalmente um erro quadrático médio pequeno se a amostra s for grande.

5º - Outros estimadores sintéticos

Os estimadores sintéticos podem assentar num modelo de regressão mais geral.

Estimador Sintético A

Suponhamos agora que o universo U está particionado em U_d domínios com dimensão N_d, $d = 1, ..., m$, isto é, $U = U_1 \cup ... \cup U_m$, $N = N_1 + ... + N_m$.
A observação de y para cada elemento do universo será indicada por y_{dk}, em que para cada domínio U_d se tem $k = 1, ..., N_d$. O modelo de regressão linear considerado será

$$y_{dk} = \boldsymbol{x_{dk}}\boldsymbol{\beta} + u_d + \varepsilon_{dk}, \tag{2.3}$$

em que $\boldsymbol{x_{dk}} = (x_{1dk}, ..., x_{qdk})$ é o vetor de q covariáveis para o elemento (dk), u_d são os efeitos aleatórios específicos do domínio U_d independentes e igualmente distribuídos com $E[u_d] = 0$ e $var[u_d] = \sigma_u^2$, e ε_{dk} são independentes entre si e de u_d com $E[\varepsilon_{dk}] = 0$ e $var[\varepsilon_{dk}] = \sigma_\varepsilon^2$. Note-se que este é um modelo misto.
Tal como anteriormente, a estimativa de y_{dk} será $E[Y_{dk}] = \boldsymbol{x_{dk}}\boldsymbol{\beta}$, visto que tanto os efeitos aleatórios como os erros têm valor médio nulo. Logo,

$$\hat{t}_{yd}^{synA} = \boldsymbol{t_{xd}}\hat{\boldsymbol{\beta}}^{unidade},$$

em que $\boldsymbol{\beta}^{unidade}$ é agora calculada pelo método da máxima verosimilhança, uma vez que se supõe que tanto os efeitos aleatórios como os erros unitários têm distribuição normal. Esta estimação é muito mais complexa do que a estimação pelo método dos mínimos quadrados em modelos simples, e obtém-se normalmente usando o algoritmo de score de Fisher (*Fisher scoring algorithm*).

Estimador Sintético B

Este estimador apoia-se no mesmo raciocínio do estimador anterior, mas é feito ao nível dos totais dos domínios que compõem o universo. O modelo de regressão, sendo estocástico, é dado por

$$t_{yd} = \boldsymbol{t_{xd}}\boldsymbol{\beta} + u_d + \zeta_d, \tag{2.4}$$

$d = 1, ..., m$ onde u_d são efeitos aleatórios e ζ_d são erros de amostragem. Supõe-se que u_d são independentes e com $E[u_d] = 0$ e $var[u_d] = \sigma_u^2$, e que ζ_d são também independentes entre si e dos u_d com $E[\zeta_d] = 0$ e $var[\zeta_d] = D_d$ (normalmente supõe-se $D_d = N_d\sigma_\varepsilon$).
Partindo de considerações semelhantes às anteriores, conclui-se que o estimador sintético B é dado por uma equação semelhante à do modelo anterior

$$\hat{t}_{yd}^{synB} = \boldsymbol{t_{xd}}\hat{\boldsymbol{\beta}}^{dominio},$$

mas em que a estimativa de $\boldsymbol{\beta}$ se faz de modo diferente do anterior, mas também de forma iterativa. Nessa iteração, D_d é normalmente substituída pela sua estimativa $\hat{D}_d = n_d\hat{\sigma}_\varepsilon$ em que $\hat{\sigma}_\varepsilon$ é, por sua vez, uma estimativa agregada de σ_ε.

6^0 - Estimadores BLUP e EBLUP

Tal como no caso anterior, os estimadores BLUP e EBLUP chamam-se A ou B consoante a estimação é feita, respetivamente, ao nível da unidade ou do domínio.

Estimadores BLUP e EBLUP A

O estimador BLUP A baseia-se no modelo 2.3, mas agora as estimativas dos totais são encontradas através da estratégia BLUP (*Best Linear Unbiased Prediction*)

$$minimizar E[\hat{t}_{yd}^{BLUPA} - t_{yd}]^2.$$

No caso de se conhecerem os parâmetros de variância σ_u^2 e σ_ϵ^2, então o estimador BLUP pode ser escrito como

$$\hat{t}_{yd}^{BLUPA} = \boldsymbol{t_{xd}}\hat{\boldsymbol{\beta}}^{unidade} + \gamma_d(\hat{t}_{yd}^{GREG} - \boldsymbol{t_{xd}}\hat{\boldsymbol{\beta}})^{unidade},$$

em que $\gamma_d = \sigma_u^2/(\sigma_u^2 + n_d\sigma_\epsilon^2)$. Este estimador ainda pode ser considerado como um estimador composto de um estimador GREG com o estimador sintético A:

$$\hat{t}_{yd}^{BLUPA} = \gamma_d\hat{t}_{yd}^{GREG} + (1-\gamma_d)\hat{t}_{yd}^{synA}.$$

No caso de não se conhecerem as variâncias do estimador, as expressões acima são válidas substituindo as variâncias pelas suas estimativas, e usando a estimativa EBLUP de β (encontrada através de *Empirical Best Linear Unbiased Prediction*). Nesse caso teremos

$$\hat{t}_{yd}^{EBLUPA} = \hat{\gamma}_d\hat{t}_{yd}^{GREG} + (1-\hat{\gamma}_d)\hat{t}_{yd}^{synA}.$$

Estimadores BLUP e EBLUP B

Para o caso em que o modelo de regressão utilizado é do tipo 2.4, ou seja, construído apenas com informação ao nível dos domínios, podemos construir os estimadores BLUP B e EBLUP B com pressupostos idênticos aos do caso anterior. O estimador BLUP B pode escrever-se sob a forma de estimador composto do estimador direto com o estimador sintético,

$$\hat{t}_{yd}^{BLUPB} = \gamma_d\hat{t}_{yd} + (1-\gamma_d)t_{yd}^{synB}.$$

Capítulo 3

Inquérito ao Emprego

3.1 Metodologia

Neste capítulo irá ser feita uma abordagem à metodologia do Inquérito ao Emprego, no que diz respeito à amostragem, estimação e avaliação da precisão das estimativas.
O Inquérito ao Emprego (IE), conduzido pelo Instituto Nacional de Estatística, teve início em 1972 e surgiu da necessidade de caracterizar a população face ao mercado de trabalho. É um inquérito contínuo, cujos apuramentos são divulgados trimestralmente e dirige-se a residentes em alojamentos familiares de residência principal, no território nacional.
O questionário permite conhecer a situação socioeconómica dos indivíduos da amostra numa determinada semana, chamada semana de referência, que geralmente é a semana anterior àquela em que os indivíduos são entrevistados presencialmente ou pelo telefone.

3.1.1 Amostragem

Dimensionamento da amostra

A dimensão da amostra foi definida em cumprimento do artigo 3º do Regulamento (CE) nº577/98, de 9 de Março de 1998, assim como de diretrizes nacionais.
Relativamente ao Regulamento, os critérios adotados foram os seguintes:

- Precisão em nível - para um grupo de pessoas desempregadas que represente 5% da população em idade ativa, o desvio-padrão relativo (ou coeficiente de variação) da estimativa das médias anuais, ao nível NUTS II, não deverá exceder 8%. As regiões com menos de 300 000 habitantes ficam isentas desta condição.

- Precisão em evolução - no caso de um inquérito contínuo, para as subpopulações que constituam 5% da população em idade ativa, o desvio-padrão da estimativa das variações entre dois trimestres consecutivos, ao nível nacional, não deverá exceder 2%. Para os Estados-Membros cuja população varie entre um milhão e vinte milhões, o desvio-padrão relativo da estimativa das variações trimestrais não deverá exceder, ao nível nacional, 3%.

Supondo tratar-se de uma amostragem aleatória simples e exigindo as condições acima mencionadas, é possível obter o número de indivíduos em idade ativa que deverão pertencer à amostra do IE. No entanto, a amostra do IE é constituída por unidades de alojamento. Para se obter uma aproximação do número de alojamentos que deverão pertencer à amostra, pode dividir-se o número pretendido de indivíduos em idade ativa pelo número médio de pessoas em idade ativa em cada alojamento.
É possível encontrar a dimensão da amostra para um esquema de amostragem complexo, efetuando um ajustamento através dos efeitos do desenho da amostra na precisão das estimativas.
Dado que as dimensões calculadas segundo o Regulamento seriam, em algumas regiões, insuficientes para permitirem a divulgação de estimativas para certas desagregações (segundo critérios nacionais), decidiu-se, para tal fim, aumentar a dimensão da amostra do IE nessas regiões.

Seleção da amostra

A amostra do IE é selecionada a partir de uma base de amostragem denominada por Amostra-Mãe. A base de amostragem em vigor é designada por AM-2001, e foi selecionada com base no Recenseamento da População e Habitação de 2001 (Censos 2001), utilizando um esquema aleatório complexo que incluiu estratificação e seleção sistemática de conglomerados com probabilidade proporcional à dimensão das áreas estatísticas em que estão incluídos.
A AM-2001 é constituída por 1408 áreas estatísticas distribuídas por 35 estratos, tendo-se listado em cada área todos os correspondentes alojamentos familiares.
A amostra do IE envolve duas etapas: na primeira são selecionadas áreas estatísticas (AM); e na segunda são selecionados alojamentos familiares. Não é necessária uma terceira etapa de seleção dos indivíduos porque são inquiridos todos os indivíduos do alojamento de residência principal.
No esquema de seleção é exigido que as unidades de alojamento tenham a mesma probabilidade de seleção (amostra auto-ponderada).

Na primeira etapa, as áreas são selecionadas sistematicamente em cada estrato com probabilidade proporcional ao número de alojamentos de residência principal. Dentro de cada área selecionada na primeira etapa, selecionam-se dois blocos sistemáticos de alojamentos sequenciais. Desta forma, consegue-se reduzir os custos de deslocação.
A amostra obedece a um esquema de rotação na transição de um trimestre para outro. Cada amostra é dividida em seis subamostras, sendo que em cada trimestre uma delas (aquela que pertence à amostra há mais tempo) é substituída por uma nova subamostra. Assim, cada conjunto de alojamentos permanece na amostra durante seis trimestres consecutivos.

3.1.2 Estimação

Para construir as estimativas é aplicado um ponderador a cada unidade estatística. Este ponderador pode ser interpretado como o peso do indivíduo, isto é, o número de indivíduos que ele representa na população. Resulta do produto de três fatores: um fator inicial, dado pelo inverso da probabilidade de seleção de cada unidade; um fator de correção para as não respostas; e um fator de calibragem da amostra, utilizando informação externa ao inquérito, por um processo denominado *ajustamento por margens*.
Considere-se uma população finita U com N indivíduos, que se subdivide em R regiões NUTS III, e da qual se retirou uma amostra s de dimensão n. Pretende estimar-se o total de y para U (Equação 2.1).

Sejam:
π_k^{-1} - inverso da probabilidade de seleção do indivíduo k
$\hat{X}_r$ - estimativa da população na região r (NUTS III) a que pertence o indivíduo k
x_r - número de inquiridos na região r (NUTS III) a que pertence o indivíduo k

O ponderador inicial (d_k) é dado por:

$$d_k = \pi_k^{-1} \frac{\hat{X}_r}{\sum_{k=1}^{x_r} \pi_k^{-1}}$$

Note-se que este ponderador já inclui uma correção para as não respostas.

Sejam $X_1, ..., X_j, ..., X_J$ J variáveis auxiliares conhecidas sobre a amostra s, e cujos totais

se conhecem para a população. O estimador do total de y usado pelo IE é dado por

$$\hat{t}_y = \sum_{k\epsilon s} w_k y_k$$

em que os pesos finais ajustados w_k são tão próximos quanto possível dos pesos iniciais d_k, mas verificando as equações de ajustamento:

$$\sum_{k\varepsilon s} w_k x_{jk} = X_j, \quad \forall\ j = 1, ..., J.$$

Estas condições garantem que as variáveis auxiliares, extrapoladas com os ponderadores ajustados, sejam iguais aos totais populacionais conhecidos.

O método de *ajustamento por margens* consiste em minimizar uma função que mede a soma ponderada das distâncias entre os ponderadores iniciais (d_k) e os ponderadores finais ajustados (w_k), sujeita às condições de ajustamento. Seja G a função de distância de argumento w_k/d_k que vai medir as distâncias entre os d_k e os w_k. Esta função deve ser positiva e convexa, e $G(1) = G'(1) = 0$.
Uma vez escolhida a função, o problema reduz-se a determinar os pesos w_k ($k\varepsilon s$) que são soluções de:

$$Min \sum_{k\varepsilon s} d_k G(w_k/d_k) \quad \text{sob as condições de ajustamento} \quad \sum_{k\varepsilon s} w_k \boldsymbol{x_k} = \boldsymbol{X}.$$

Introduzindo um vetor de multiplicadores de Lagrange $\boldsymbol{\lambda'} = (\lambda_1, ..., \lambda_J)$, vem

$$L = \sum_{k\varepsilon s} d_k G(w_k/d_k) - \boldsymbol{\lambda'} \left(\sum_{k\varepsilon s} w_k x_k - \boldsymbol{X} \right).$$

Para minimizar a função em causa, deriva-se L e iguala-se a zero, resultando

$$w_k = d_k F(x_k' \boldsymbol{\lambda})$$

em que F é a função inversa da derivada da função G. O vetor $\boldsymbol{\lambda}$ é calculado mediante a resolução de um sistema não linear, que se pode resolver recorrendo ao método iterativo de Newton.
A função de distância $G(x)$ (em que $x = w_k/d_k$) utilizada no IE é a do *método logit (método ranking ratio truncado)*, que é dada por:

$$G(x) = \left((x - L) Log \frac{x - L}{1 - L} + (U - x) Log \frac{U - x}{U - 1} \right) \frac{1}{A}, \quad \text{se} \ \ L < x < U$$

em que $A = \frac{U-L}{(1-L)(U-1)}$ e $F(u) = \frac{L(U-1)+U(1-L)exp(Au)}{U-1+(1-L)exp(Au)}\ \epsilon\ ]L, U[$ com $u = x_k'\lambda$.

Note-se que as relações entre os pesos ajustados e os pesos iniciais (w_k/d_k) estão limitadas inferiormente por L e superiormente por U.

A partir da série iniciada no 1º trimestre de 2011, as margens utilizadas foram as seguintes:

- população por região NUTS II (NUTS - 2002), sexo e escalão etário (0 a 4 anos; ... ; 10 a 14 anos; 15 a 17 anos; ...; 70 a 74 anos; 75 e mais anos);
- população por região NUTS III ou agregações de regiões NUTS III e sexo;
- população por região NUTS III ou agregações de regiões NUTS III e escalão etário (0 a 14 anos; 15 a 24 anos; 25 a 34 anos; 35 a 44 anos; 45 a 64 anos; 65 e mais anos).

3.1.3 Precisão

Como é evidente, as estimativas obtidas estão sujeitas a erros. A precisão de um estimador pode ser medida em termos abslolutos (variância e desvio-padrão) ou em termos relativos (coeficiente de variação).
A variância deste tipo de estimadores tem uma expressão difícil de obter explicitamente e, por isso, recorre-se a métodos numéricos para o seu cálculo. No IE, o método utilizado é o Jackknife. Passa a fazer-se uma descrição muito resumidamente deste método.
Considere-se $x_1, x_2, ..., x_n$ uma amostra com n elementos, que se pode dividir em g grupos. Seja θ o parâmetro de interesse e $\hat{\theta}$ um estimador desse parâmetro. Pretende-se estimar a variância do estimador de θ. A ideia que está subjacente ao método é considerar que a amostra é todo o universo. Pode descrever-se como um método de reamostragem que consiste em constituir, a partir da amostra original, g subamostras, designadas por réplicas, constituídas por $g-1$ grupos, sendo esta amostragem feita sem reposição. Ou seja, cada subamostra contém todos os grupos da amostra original, excepto um, que é escolhido sequencialmente. Na aplicação deste método ao IE, estes grupos correspondem às áreas estatísticas. Seja $\hat{\theta}_\alpha$ o estimador de θ da réplica α. O estimador da variância é dado por:

$$\widehat{var}(\hat{\theta}) = \frac{g-1}{g}\sum_{\alpha=1}^{g}(\hat{\theta}_\alpha - \hat{\theta})^2.$$

O estimador do coeficiente de variação (cv) de $\hat{\theta}$ é

$$\widehat{cv}(\hat{\theta}) = \frac{\sqrt{\hat{var}(\hat{\theta})}}{\hat{\theta}} \times 100\%.$$

O INE divulga estimativas trimestrais e anuais da taxa de desemprego ao nível NUTS II. Usando o mesmo método de estimação, não é possível obter estimativas que sejam consi-

deradas com precisão aceitável ($cv < 20\%$) ao nível NUTS III. Uma vez que o INE opta por não divulgar estimativas com coeficientes de variação elevados, surge a preocupação em estudar o comportamento de estimadores indiretos no IE. De facto, a teoria em que estes estimadores assentam indica que é possível obter ganhos de precisão nas regiões de menor dimensão.

3.2 Conceitos

Apresentam-se seguidamente os conceitos que serão usados ao longo deste trabalho tal como estão definidos no regulamento europeu do inquérito ao trabalho.

População Desempregada

Define-se população desempregada como o conjunto de indivíduos com idade dos 15 aos 74 anos que, no período de referência, se encontravam simultaneamente nas seguintes situações:

- não tinham trabalho remunerado nem qualquer outro;
- estavam disponíveis para trabalhar num trabalho remunerado ou não;
- tinham procurado um trabalho, isto é, tinham feito diligências ao longo de um período especificado (período de referência ou nas três semanas anteriores) para encontrar um emprego remunerado ou não.

População Ativa

Define-se população ativa como o conjunto de indivíduos com idade mínima de 15 anos que, no período de referência, constituiam a mão de obra disponível para a produção de bens e serviços que entram no circuito económico (empregados e desempregados).

Taxa de Desemprego

A taxa de desemprego define o peso da população desempregada sobre o total da população ativa. T.D.(%) = (População desempregada)/(População ativa) x 100%.

NUTS

NUTS é a nomenclatura das unidades territoriais para fins estatísticos, versão de 2002, estabelecida pelo decreto-lei nº244/2002 e pelo regulamento comunitário nº1059/2003

(NUTS-2002).

- Nível I - Portugal Continental, Região Autónoma dos Açores, Região Autónoma da Madeira.
- Nível II - Norte, Centro, Lisboa, Alentejo, Algarve, Região Autónoma dos Açores, Região Autónoma da Madeira.
- Nível III - Norte: Minho Lima, Cávado, Ave, Grande Porto, Tâmega, Entre Douro e Vouga, Douro, Alto Trás-Os-Montes; Centro: Baixo Vouga, Baixo Mondego, Pinhal Litoral, Pinhal Interior Norte, Pinhal Interior Sul, Dão Lafões, Serra da Estrela, Beira Interior Norte, Beira Interior Sul, Cova Beira, Oeste, Médio Tejo; Lisboa: Grande Lisboa, Península de Setúbal; Alentejo: Alentejo Litoral, Alto Alentejo, Alentejo Central, Baixo Alentejo, Lezíria Tejo; Algarve; Região Autónoma dos Açores; Região Autónoma da Madeira.

A Tabela 6.1 em anexo mostra como as NUTS III se podem agrupar em NUTS II.

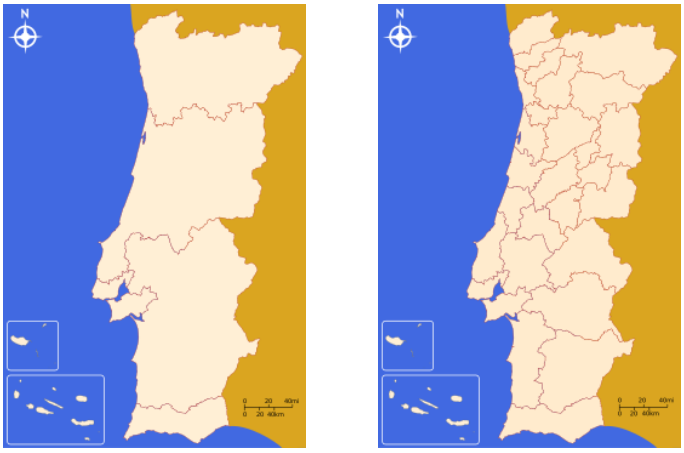

Figura 3.1: NUTS II e NUTS III de Portugal Continental

Capítulo 4

Modelo para a Estimação

Neste capítulo será apresentada uma metodologia usada pelo ONS para a estimação do desemprego para pequenas áreas, que vai ao encontro do modelo que nos parece mais razoável para a estimação ao nível NUTS III. Será aplicada essa metodologia aos dados do INE, com as devidas adaptações.

4.1 Metodologia seguida pelo ONS

O ONS tem feito um esforço, desde 1999, para desenvolver e melhorar uma metodologia de estimação em pequenos domínios a fim de fornecer estimativas do desemprego baseadas nos dados anuais do inquérito *Labour Force Survey* (LFS).
O LFS é um inquérito contínuo, de larga escala, com uma amostra de aproximadamente 53000 alojamentos em cada trimestre. É um inquérito rotativo em que os respondentes são inquiridos em cinco trimestres consecutivos.
Este inquérito é a fonte da informação nacional do mercado de trabalho. Foi primeiramente desenhado para produzir estimativas ao nível da *Government Office Region* (GOR), sendo que a amostra não tem tamanho suficiente para fornecer estimativas diretas do desemprego com precisão adequada para cada autoridade local (LA) ou constituição parlamentar (PC) na Grã-Bretanha. Contudo, foram utilizadas técnicas de estimação em pequenas áreas para resolver o problema de pequenas dimensões amostrais, e o uso destas técnicas tem permitido que a qualidade das estatísticas seja melhorada.
Atualmente, são produzidas estimativas para as LAs com base no *Annual Population Survey* (APS), que combina 1/5 de cada trimestre do LFS aumentando assim a amostra para 96000 alojamentos por ano. Estas estimativas são produzidas trimestralmente e baseiam-se nos dados anuais. Foram produzidas pelo ONS pela primeira vez em 2003, sendo publicadas como *estatísticas experimentais* e depois como *estatísticas nacionais* em

Julho de 2006.
A metodologia model-based usada pelo ONS baseia-se na determinação de uma forte relação entre o desemprego (tal como o APS mediu) e informação auxiliar. A principal fonte de informação auxiliar é o número de beneficiários do *Jobseeker's Allowance* (correspondente aos Centros de Emprego). Como esta informação provém de um sistema administrativo, os dados estão disponíveis sem erro amostral e podem ser divididos, por exemplo, em diferentes categorias de idade e sexo, bem como em unidades geográficas de nível mais pequeno. Esta relação é determinada com base num modelo de regressão, que é usado para produzir estimativas model-based para estes domínios pequenos.

[***Modelos de regressão***]
Os modelos de regressão estabelecem relações entre uma variável dependente (ou variável resposta) e uma ou mais variáveis explicativas, e têm como objetivo estimar e/ou prever o valor médio da variável dependente sabendo os valores das variáveis explicativas.
Dentro dos modelos de regressão existem os modelos de regressão generalizada. Estes modelos permitem tratar variáveis dependentes que não têm distribuição Normal, através de uma tranformação que é feita ao valor médio, permitindo assim que este seja modelado por uma função linear das covariáveis. Essa tranformação é feita através de uma função de ligação, que depende do tipo de resposta e do estudo que se está a efetuar.
Neste estudo, a variável dependente toma o valor 1 se o indivíduo está desempregado e toma o valor 0 se está empregado, logo o valor médio corresponde à probabilidade de o indivíduo estar desempregado. Quando os dados relativos à variável dependente são deste tipo (binários), é geralmente aplicada a função de ligação *logit* e, neste caso, estamos perante um *modelo logístico*. [3]
Alguns conjuntos de dados estão estruturados hierarquicamente em diversos níveis. Os modelos que apresentam mais do que um nível são chamados de modelos multiníveis, e têm em conta determinados contextos, ou níveis, em que os indivíduos estão inseridos. Neste estudo, em particular, são considerados dois níveis, o grupo resultante do cruzamento do sexo com o escalão etário (que passaremos a denotar por sexo x escalão etário) e a PC a que o indivíduo pertence.
Por vezes, a compreensão de um fenómeno depende não só das características dos indivíduos como também da sua organização. É de esperar que indivíduos inseridos num certo contexto tenham comportamentos mais semelhantes entre si do que com indivíduos num outro contexto. Nesta situação, recorre-se geralmente aos modelos mistos. Estes modelos contêm efeitos fixos e efeitos aleatórios. Os efeitos fixos representam a parte determinística da relação, enquanto que os efeitos aleatórios capturam variações e heterogeneidade associada aos indivíduos.

A metodologia usada para as LAs tem sido extendida para produzir estimativas do desemprego para as PCs, mas foi necessário fazer uma alteração porque alguns parâmetros usados no modelo original não estavam disponíveis para estas regiões.
No modelo LA entram os dados do APS para 406 LAs por grupo sexo x escalão etário (16 a 24; 25 a 49; 50 e mais). A metodologia envolve o número de desempregados na amostra, o tamanho da amostra, as estimativas diretas do total de desempregados, o total dos indivíduos ativos, e o total da população. Também é fornecido o número de inscritos do *Jobseeker's Allowance* para cada categoria, as *Government Office Region* (GOR), e a classificação sócio-económica das LAs.
O modelo construído para estimar a população desempregada ao nível das PCs entra com os mesmos dados do anterior, mas para 628 PCs, e com a exceção da variável de classificação sócio-económica. A classificação para PCs não tem sido publicada e, depois de se fazer alguma investigação, foi decidido que esse termo pode ser omitido.
Atualmente, têm sido produzidas estimativas model-based, e respetivos intervalos de confiança, da taxa de desemprego por PC, para a Inglaterra, Escócia e Países de Gales.
Estas estimativas são anuais, isto é, usam uma média dos últimos doze meses dos totais de desempregados registados e doze meses dos dados do inquérito. São publicadas todos os trimestres. Desta forma, as estimativas publicadas num dado trimestre estão fortemente correlacionadas com as estimativas do trimestre anterior porque 3/4 dos dados são comuns. Portanto, não é válido comparar estimativas entre períodos com menos de 12 meses de intervalo.
Como vimos, a forma do modelo de regressão depende da natureza da variável de interesse. Neste caso, a variável é binária e indica se um indivíduo que pertence a um dado grupo sexo x escalão etário e PC está desempregado ou não.
O modelo usado pelo ONS que relaciona a probabilidade de estar desempregado para o grupo sexo x escalão etário i ($i = 1, ..., 6$) na PC j ($j = 1, ..., 628$) com as variáveis auxiliares é dado por

$$logit(p_{ij}) = \beta_0 + \beta_i I_{see(i)} + \beta_{GOR(k)} I_{GOR(k)} + \beta_{17} X_{ij} + \beta_{18} X_j + \beta_{cc(i)} X_{ij} I_{see(i)} + u_j$$

onde

$$logit(p_{ij}) = log_e \left(\frac{p_{ij}}{1 - p_{ij}} \right)$$

e

- $I_{see(i)}$ e $I_{GOR(k)}$ são indicadores do grupo sexo x escalão etário i e da GOR k (onde a PC j pertence), com $k = 1, ..., 11$;
- X_{ij} e X_j são o logit da proporção de desemprego registado em cada grupo sexo x escalão etário na PC e o logit da proporção de desemprego registado na PC, respetivamente;

- $X_{ij}I_{see(i)}$ representa a interação entre o desemprego registado e o grupo sexo x escalão etário;
- u_j é o efeito aleatório da área;
- $\beta_0, \beta_i, \beta_{GOR(k)}, \beta_{17}, \beta_{18}, \beta_{cc(i)}$ são os coeficientes de regressão com $i = 1, ..., 5$ e $k = 1, ..., 11$ porque um grupo sexo x escalão etário e uma GOR corresponde à categoria *baseline*.

Como podemos ver, o modelo usado para a estimação em pequenos domínios ao nível das PCs é um modelo logístico misto multinível e com intercept aleatório. Os efeitos aleatórios das PCs "capturam"as fontes de variação e a heterogeneidade da área que não são explicadas pelos dados auxiliares.

Note-se que o modelo dá uma estimativa para a proporção de desemprego dada por nº desempregados/população com 15 ou mais anos. A estimativa model-based para o total de desempregados D em cada PC é dada pela combinação do número de desempregados na amostra n e a estimativa do número de desempregados na componente não amostral (complementar) obtida pelo modelo, ou seja, $\hat{D} = n + \hat{p} * (N - n)$ onde N é a dimensão da população.

Para assegurar que as estimativas model-based são consistentes com as estimativas publicadas para níveis geográficos maiores, as estimativas model-based são calibradas pelas estimativas diretas APS do desemprego para as GORs em Inglaterra, Países de Gales, e Escócia. Além disso, as estimativas para PCs que correspondem à mesma área geográfica das LAs, são calibradas de forma a serem iguais. Aditivamente, quando um determinado número de PCs corresponde a uma LA, o total estimado de desempregados é calibrado pela estimativa para a LA.

Como vimos, a taxa de desemprego é o quociente entre o número de desempregados e o número de pessoas ativas. O ONS estima o número de pessoas ativas através da soma da estimativa model-based do total de desempregados com a estimativa direta do número de empregados.

Através dos gráficos que comparam os coeficientes de variação ($cv's$) das estimativas model-based com os $cv's$ das estimativas diretas, que são apresentados no artigo [3], é possível verificar que as estimativas model-based são mais precisas. Contudo, apesar de serem mais precisas, estas estimativas são enviezadas. O objetivo é obter estimativas com boa precisão e com um bias tão pequeno quanto possível.

No mesmo artigo é apresentado um gráfico que relaciona os dois tipos de estimativas, sendo que se aproximam da reta $y = x$, o que permite concluir que se as estimativas diretas forem centradas então as estimativas model-based serão no mínimo aproximadamente centradas. Como foi mostrado anteriormente, as estimativas diretas são centradas.

4.1.1 Aplicação da metodologia aos dados do INE

Com a aplicação da metodologia exposta aos dados do INE, irão ser produzidas estimativas trimestrais (para o $1^{\underline{o}}$ trimestre de 2011), e estimativas anuais (para o ano 2011, do $1^{\underline{o}}$ trimestre ao $4^{\underline{o}}$ trimestre). A amostra anual pode ser selecionada a partir das amostras dos 4 trimestres que compõem o ano que termina nesse trimestre, sendo necessário evitar a dependência entre elas. Como já foi referido, cada amostra trimestral é constituída por 6 subamostras, havendo rotatividade destas de um trimestre para outro. Nesta transição, a subamostra que está há mais tempo é substituída por uma nova subamostra. Desta forma, a amostra de um trimestre tem em comum 5/6 com o trimestre consecutivo. Se selecionarmos os valores correspondentes à rotação de entrada e à rotação de saída em cada trimestre que compõe o ano em causa, ficamos com uma amostra com 8 subamostras independentes. Como podemos ver no esquema representado na Figura 4.1, sobra ainda uma rotação que não é considerada e que podemos acrescentar a esta amostra de forma a aumentar o seu tamanho, mantendo a independência entre as observações. As observações desta rotação podem ser selecionadas de qualquer um dos quatro trimestres. No entanto, em vez disso, pode-se subdividir a rotação em quartos para recolher as observações de cada um deles em trimestres diferentes. Desta forma, evitamos um enviesamento temporal das observações. A esta divisão, deve ainda ser exigido que indivíduos selecionados se distribuam da mesma forma pelas 1408 áreas da amostra e que, além disso, indivíduos pertencentes ao mesmo agregado sejam selecionados no mesmo trimestre. Não é possível satisfazer todas estas condições simultaneamente, no entanto, para tentar optimizar a solução, optou-se por distribuir as áreas sequencialmente pelos quatro trimestres.

Constituídas as amostras, interessa-nos aplicar o modelo do ONS, com as devidas adaptações. Para isso, é necessário o conhecimento de informação auxiliar.
Existem várias fontes possíveis de informação auxiliar, como por exemplo, o IEFP (Instituto do Emprego e Formação Profissional), o Ministério do Trabalho, a Segurança Social, etc.
Depois de um estudo sobre os dados que estas fontes podem disponibilizar, concluiu-se que há mais vantagens em recorrer à informação proveniente do IEFP.
O IEFP fornece informação do número de desempregados registados nos Centros de Emprego por grupo resultante do cruzamento do sexo com o escalão etário, ao nível do município. Esta informação, completamente independente do Inquérito ao Emprego, pode ser muito útil para as estimativas pretendidas, por se encontrar desagregada até ao nível que necessitamos e, portanto, capaz de ser incluída num modelo de regressão.
A informação do número de inquiridos e do número de desempregados na amostra do IE é agregada em 224 grupos (cruzamento entre 8 grupos sexo x escalão etário (15 a 24, 25

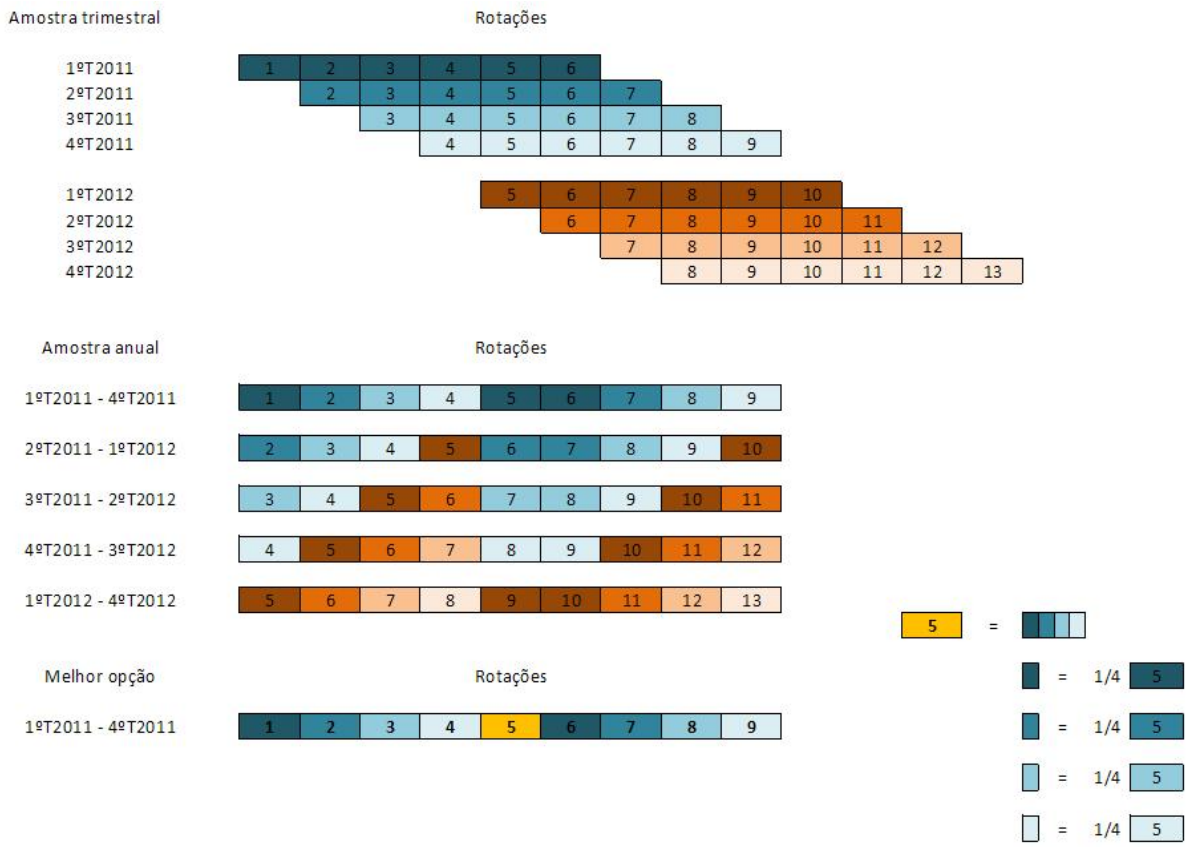

Figura 4.1: Esquema de rotações

a 34, 35 a 49, 50 e mais anos) e 28 NUTS III (referentes a Portugal Continental).
Relativamente à informação proveniente do IEFP, se pretendermos calcular estimativas trimestrais é feita uma média dos três meses do trimestre em causa, e se pretendermos obter estimativas anuais é feita uma média dos doze meses do ano em questão. Esta informação também é agregada em 224 grupos.
Quando feita a aplicação do modelo apresentado aos dados do IE, verificou-se que os cruzamentos entre a proporção de inscritos nos Centros de Emprego e os grupos sexo x escalão etário não são significativos. Além disso, a variável proporção de desemprego registado por NUTS III também não é significativa para o modelo, não tendo por isso incluído estas variáveis no estudo feito.

Considere-se então o modelo que relaciona a probabilidade de desemprego p_{ij} para um grupo sexo x escalão etário i $(i = 1, ..., 8)$ na NUTS III j $(j = 1, ..., 28)$ com as variáveis auxiliares:

$$logit(p_{ij}) = \beta_0 + \beta_i I_{idade.sexo(i)} + \beta_{NUTSII(k)} I_{NUTSII(k)} + \beta_{36} X_{ij} + u_j$$

onde

$$logit(p_{ij}) = log_e\left(\frac{p_{ij}}{1-p_{ij}}\right)$$

e

- $I_{idade.sexo(i)}$ e $I_{NUTSII(k)}$ são indicadores *dummy* do grupo sexo x escalão etário i e da região NUTS II k (onde as NUTS III se inserem);
- X_{ij} é o logit da proporção de desemprego registado em cada grupo sexo x escalão etário i na NUTS III j;
- u_j é o efeito aleatório da área; e
- $\beta_0, \beta_i, \beta_{NUTSII(k)}, \beta_{35}$ são os coeficientes de regressão com $i = 1, ..., 7$ e $k = 1, ..., 27$ porque um grupo sexo x escalão etário (masculino com menos de 25 anos) e uma região NUTS II correspondem a categorias *baseline*.

Repare-se que este estudo foi feito apenas para 28 NUTS III. Há três regiões NUTS III que coincidem com as NUTS II: Algarve, Região Autónoma da Madeira, e Região Autónoma dos Açores. Para estas regiões existem estimativas obtidas pelo método direto com precisão aceitável, tornando-se assim desnecessário incluí-las no modelo. No entanto, se houver trocas significativas de emprego entre estas regiões e as restantes NUTS III, devem ser incluídas no modelo, uma vez que influencia as estimativas obtidas para as outras NUTS III. Espera-se que estas trocas sejam significativas entre o Algarve e as restantes NUTS III de Portugal Continental, e por esse motivo esta região foi incluída no modelo. Relativamente às regiões autónomas, estas não foram incluídas por dois motivos: não temos informação auxiliar desagregada pelos grupos sexo x escalão etário referidos e, além disso, acreditamos que as trocas de emprego não são significativas.
O modelo descrito foi aplicado usando o software R (ver anexos). A função usada foi *glmer*, visto tratar-se de um modelo linear generalizado com efeitos aleatórios.
Note-se que o modelo não dá diretamente uma estimativa para a taxa de desemprego, mas sim para uma proporção de desemprego dada por: nº desempregados/população com 15 ou mais anos.
Comecemos por estimar o total de desempregados. Seja $\hat{p}_{ji}$ a estimativa da proporção obtida pelo modelo, para o grupo sexo x escalão etário i da NUTS III j. Para estimarmos o total de desempregados no grupo i da NUTS III j, t_{yji}, fazemos $\hat{t}_{yji} = \hat{p}_{ji} * N_{ji}$, em que N_{ji} é a estimativa do total de indivíduos com 15 ou mais anos na população no grupo i da NUTS III j.
Se, num futuro próximo, o INE publicar estimativas produzidas ao nível NUTS III, estas devem ser consistentes com as estimativas publicadas por NUTS II, ou seja, a soma das estimativas dos totais em todas as NUTS III pertencentes a uma NUTS II deve coincidir

com a estimativa do total nessa NUTS II. Além disso, como as estimativas ao nível NUTS II são publicadas por grupo sexo x escalão etário, também devemos exigir consistência a este nível.

As estimativas obtidas diretamente do modelo não são consistentes com as estimativas publicadas ao nível NUTS II. No entanto, podem ser ajustadas de forma a assegurar a consistência.

Seja $\hat{Y}_p^{INE}$ o estimador do total Y_p na NUTS II p. Assuma-se que a NUTS II p é uma partição de D_p NUTS III. Sejam $\hat{Y}_1, ..., \hat{Y}_{D_p}$ os estimadores dos totais $Y_1, ..., Y_{D_p}$.

Em geral, a propriedade de consistência $\hat{Y}_p^{INE} = \sum_{d=1}^{D_p} \hat{Y}_d$ não é satisfeita. Nestes casos, $\hat{Y}_1, ..., \hat{Y}_{D_p}$ podem ser transformados em estimadores consistentes pelo seguinte cálculo [6]:

$$\hat{Y}_d^c = \lambda_{yp} \hat{Y}_d$$

onde

$$\lambda_{yp} = \frac{\hat{Y}_p^{INE}}{\sum_{d=1}^{D_p} \hat{Y}_d}.$$

Estes estimadores satisfazem $\hat{Y}_p^{INE} = \sum_{d=1}^{D_p} \hat{Y}_d^c$.

Se aplicarmos esta calibração a cada um dos grupos sexo x escalão etário, então obtemos estimativas para a população desempregada por grupo nas NUTS III, consistentes com os grupos nas NUTS II, que posteriormente serão agregadas por NUTS III. Como é de esperar, se agregarmos estas estimativas por NUTS II, obtemos estimativas iguais às estimativas diretas que são divulgadas a este nível.

Capítulo 5

Análise dos Resultados

Neste capítulo serão apresentados os resultados obtidos a partir do modelo descrito no capítulo anterior e calculadas as estimativas para os totais e taxa de desemprego ao nível NUTS III, para um trimestre (1^{o} trimestre de 2011) e para um ano (2011).
Vai ser feita ainda uma abordagem a alguns indicadores de performance, que serão estimados com base em métodos numéricos, para que possamos avaliar a qualidade das estimativas obtidas.

5.1 Estimativas

Nas Tabelas 5.1 e 5.2 podemos ver as estimativas obtidas para a população desempregada por NUTS II e por NUTS III pelo método direto e pelo método descrito no capítulo anterior, antes da calibração e depois da calibração, para o 1^{o} trimestre de 2011 e para o ano completo 2011, respetivamente.
Note-se que as estimativas calibradas não diferem muito das estimativas obtidas antes da calibração, no entanto, pode reparar-se que as estimativas obtidas a partir do modelo apresentado são, para a maior parte das NUTS III, inferiores às estimativas diretas. A razão poderá estar no efeito que os dados do IEFP têm sobre as estimativas model-based obtidas e, de facto, verificou-se que para a maior parte das NUTS III a população inscrita nos CE é inferior às estimativas diretas da população desempregada.
Nos resultados que se seguem considere-se que as estimativas model-based são as obtidas após calibração.

Os gráficos representados na Figura 5.1 permitem concluir que as estimativas model-based obtidas por NUTS III são próximas das estimativas diretas para o mesmo nível, pois distribuem-se praticamente sobre a reta $y = x$. Como as estimativas diretas são

NUTS II	NUTS III	População desempregada		
			Modelo	
		Direta	Antes da calibração	Depois da calibração
Norte		**254531.8**	**234129.0**	**254531.8**
	Minho-Lima	16350.3	13037.9	14174.0
	Cávado	27581.7	24440.7	26623.4
	Ave	38194.8	33819.5	36675.9
	Grande Porto	99176.5	90265.6	97725.6
	Tâmega	37437.3	34995.8	38436.8
	Entre Douro e Vouga	13651.7	14769.9	15992.1
	Douro	12185.1	11912.9	13007.8
	Alto Trás-os-Montes	9954.3	10886.7	11896.2
Centro		**124171.2**	**115649.9**	**124171.2**
	Baixo Vouga	25313.5	20655.4	22265.0
	Baixo Mondego	17137.1	15208.4	16252.0
	Pinhal Litoral	13080.9	11529.7	12399.0
	Pinhal Interior Norte	5584.6	5861.4	6315.1
	Pinhal Interior Sul	2801.4	1484.9	1593.9
	Dão-Lafões	7242.6	12493.8	13418.3
	Serra da Estrela	2201.7	2421.8	2579.9
	Beira Interior Norte	5348.1	4813.5	5154.0
	Beira Interior Sul	1438.4	3451.8	3669.9
	Cova da Beira	4045.8	4840.7	5142.8
	Oeste	26442.2	22173.0	23836.8
	Médio Tejo	13534.9	10715.5	11544.5
Lisboa		**195392.2**	**174855.5**	**195392.2**
	Grande Lisboa	137192.8	122201.3	136468.2
	Península de Setúbal	58199.4	52654.1	58924.0
Alentejo		**46524.9**	**42646.2**	**46524.9**
	Alentejo Litoral	4024.9	4564.6	4972.1
	Alto Alentejo	11677.8	8208.7	9006.5
	Alentejo Central	8469.5	8551.1	9388.4
	Baixo Alentejo	7901.3	7590.0	8265.5
	Lezíria do Tejo	14451.3	13731.8	14892.3
Algarve		**38628.7**	**34749.6**	**38628.7**

Tabela 5.1: Estimativas da população desempregada, pelo método direto e pelo modelo (antes e depois da calibração), por NUTS II e NUTS III de Portugal Continental, referentes ao 1º trimestre de 2011

centradas, esta relação significa que as estimativas model-based obtidas são aproximadamente centradas. Depois de obtidas as estimativas finais para a população desempregada por NUTS III, podemos calcular as estimativas para a taxa de desemprego a esse nível, Tabela 5.3. Estas estimativas da taxa de desemprego foram calculadas com base na mesma metodologia do ONS.

Mais uma vez, devido à calibração feita, as estimativas produzidas ao nível NUTS II coincidem com as estimativas diretas oficiais.

Note-se que, segundo as estimativas produzidas, o Algarve é a região NUTS II que apresenta uma maior taxa de desemprego (17.0% no 1ºtrimestre de 2011 e 15.1% em 2011), seguida de Lisboa (13.6% no 1ºtrimestre de 2011 e 14.2% em 2011). O Centro é a região

NUTS II	NUTS III	População desempregada		
			Modelo	
		Direta	Antes da calibração	Depois da calibração
Norte		**278836.7**	**263441.7**	**278836.7**
	Minho-Lima	13532.1	14595.0	15407.8
	Cávado	31518.5	25640.3	27111.8
	Ave	42626.2	40282.1	42698.2
	Grande Porto	104916.4	101252.1	107219.5
	Tâmega	45927.0	39137.2	41476.4
	Entre Douro e Vouga	17465.1	18053.3	19101.1
	Douro	13917.1	12717.0	13425.4
	Alto Trás-os-Montes	8934.2	11764.7	12396.4
Centro		**129550.5**	**133233.3**	**129550.5**
	Baixo Vouga	21922.4	23578.3	22958.3
	Baixo Mondego	16643.3	17420.1	16951.0
	Pinhal Litoral	14399.0	13766.7	13391.2
	Pinhal Interior Norte	5213.3	7030.0	6836.2
	Pinhal Interior Sul	1440.6	1770.3	1720.2
	Dão-Lafões	17674.7	15046.4	14637.1
	Serra da Estrela	2350.4	2641.1	2569.6
	Beira Interior Norte	3388.0	5240.3	5104.4
	Beira Interior Sul	2549.5	3726.1	3634.9
	Cova da Beira	3612.9	5188.3	5048.7
	Oeste	23620.0	24530.2	23802.9
	Médio Tejo	16736.3	13295.5	12895.9
Lisboa		**204754.9**	**199558.9**	**204754.9**
	Grande Lisboa	145355.9	140473.1	144090.5
	Península de Setúbal	59399.0	59085.8	60664.4
Alentejo		**45694.8**	**44643.8**	**45694.8**
	Alentejo Litoral	4234.6	4851.1	4963.5
	Alto Alentejo	10312.3	7771.7	7950.2
	Alentejo Central	7581.3	8961.8	9172.3
	Baixo Alentejo	8412.8	7939.5	8123.1
	Lezíria do Tejo	15153.8	15119.7	15485.7
Algarve		**34995.0**	**34332.4**	**34995.0**

Tabela 5.2: Estimativas da população desempregada, pelo método direto e pelo modelo (antes e depois da calibração), por NUTS II e NUTS III de Portugal Continental, referentes a 2011 (do 1ºtrimestre ao 4ºtrimestre)

NUTS II que tem menor taxa de desemprego (9.7% no 1ºtrimestre de 2011 e 10.1% em 2011). Podemos notar ainda que dentro da NUTS II Norte, o Grande Porto é a região que apresenta uma maior taxa; relativamente ao Centro, a região com maior taxa é a Beira Interior Sul para o 1º trimestre de 2011 (12.9%) e é Oeste para o ano 2011 (12.3%); em Lisboa, é a Península de Setúbal que apresenta maior valor (14.5% no 1ºtrimestre de 2011 e 15.0% em 2011); o Alto Alentejo é a região do Alentejo com maior taxa (18.2% no 1ºtrimestre e 16.0% em 2011).

[***Nota***] Nos gráficos a seguir apresentados, as NUTS III estão codificadas de 1 a 28. Estes valores correspondem à ordem pela qual estão apresentadas as NUTS III na tabela em anexo.

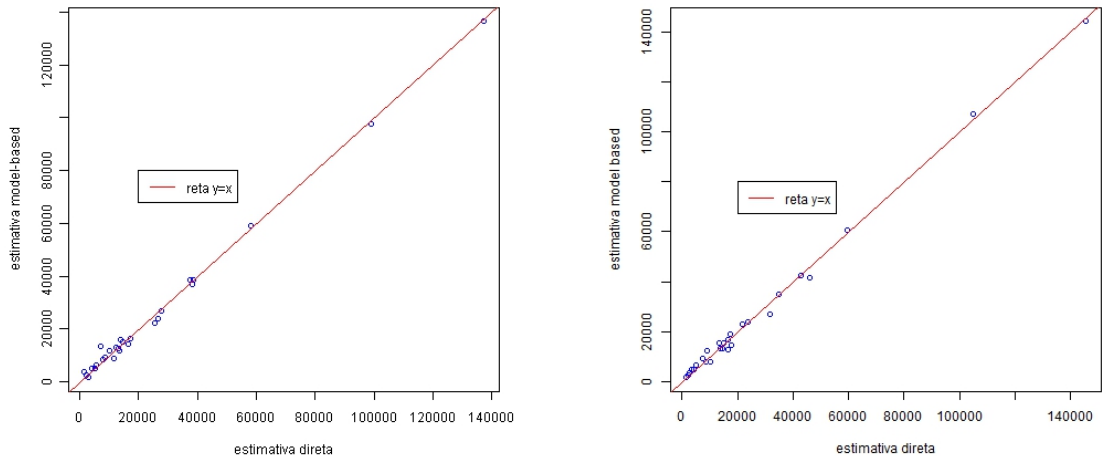

Figura 5.1: Estimativas diretas do total de desempregados versus estimativas model-based do total de desempregados, por NUTS III, para o 1º trimestre de 2011 (à esquerda) e para o ano 2011 (à direita)

Pelos gráficos representados na Figura 5.2, vê-se facilmente que a região NUTS III 25 (Alto Alentejo) é aquela que tem maior taxa de desemprego. Já as regiões que apresentam taxas de desemprego mais baixas são as regiões 13 (Pinhal Interior Sul) e 16 (Beira Interior Norte).

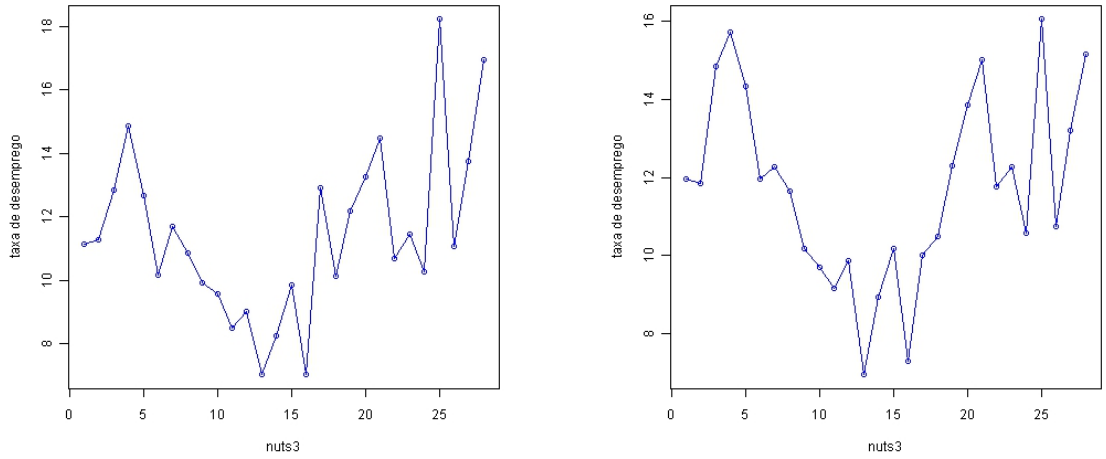

Figura 5.2: Estimativa da taxa de desemprego por NUTS III, relativa ao 1º trimestre de 2011 (à esquerda) e ao ano 2011 (à direita)

Os mapas representados na Figura 5.3 dão uma ideia da distribuição geográfica das estimativas obtidas. Note-se que grande parte das regiões com menor taxa de desemprego pertencem ao interior, o que pode contrariar a nossa intuição. No entanto, é de ter em conta que estas regiões têm uma população mais envelhecida.

Tal como foi descrito nesta secção, são necessários vários procedimentos para obtermos estas estimativas. O esquema representado na Figura 5.4 resume a informação necessária,

NUTS II	NUTS III	Taxa de Desemprego (%) 1ºT-2011	2011
Norte		**12.8**	**14.0**
	Minho-Lima	11.1	12.0
	Cávado	11.3	11.8
	Ave	12.8	14.8
	Grande Porto	14.8	15.7
	Tâmega	12.7	14.3
	Entre Douro e Vouga	10.2	12.0
	Douro	11.7	12.3
	Alto Trás-os-Montes	10.9	11.6
Centro		**9.7**	**10.1**
	Baixo Vouga	9.9	10.2
	Baixo Mondego	9.6	9.7
	Pinhal Litoral	8.5	9.2
	Pinhal Interior Norte	9.0	9.9
	Pinhal Interior Sul	7.0	7.0
	Dão-Lafões	8.3	8.9
	Serra da Estrela	9.8	10.2
	Beira Interior Norte	7.0	7.3
	Beira Interior Sul	12.9	10.0
	Cova da Beira	10.1	10.5
	Oeste	12.2	12.3
	Médio Tejo	10.7	11.8
Lisboa		**13.6**	**14.2**
	Grande Lisboa	13.3	13.9
	Península de Setúbal	14.5	15.0
Alentejo		**12.5**	**15.1**
	Alentejo Litoral	10.3	10.6
	Alto Alentejo	18.2	16.0
	Alentejo Central	11.1	10.8
	Baixo Alentejo	13.8	13.2
	Lezíria do Tejo	11.5	12.3
Algarve		**17.0**	**15.1**

Tabela 5.3: Estimativas da taxa de desemprego (em %) por NUTS II e NUTS III de Portugal Continental, referentes ao 1º trimestre de 2011 e ao ano 2011 (do 1ºtrimestre ao 4ºtrimestre)

em termos práticos. Em suma, no modelo logit apresentado entram os totais de desempregados na amostra do IE e os totais de desempregados registados nos Centros de Emprego, por grupo sexo x escalão etário x NUTS III. Este modelo fornece a estimativa para a proporção de desempregados, que é dada pelo total desempregados/total população com 15 ou mais anos, para cada um dos 224 grupos. Para estimar a população desempregada, multiplicou-se a proporção obtida, pela estimativa do total da população com 15 ou mais anos. A consistência com as estimativas oficiais é garantida através de uma calibração, a partir das estimativas diretas por grupo sexo x escalão etário e NUTS II. O objetivo é produzir estimativas do desemprego por NUTS III, sendo feita portanto uma agregação das estimativas produzidas por grupo sexo x escalão etário x NUTS III a este nível. Por fim, a taxa de desemprego é calculada a partir da estimativa da população desempregada

e recorrendo à estimativa direta da população empregada por NUTS III.

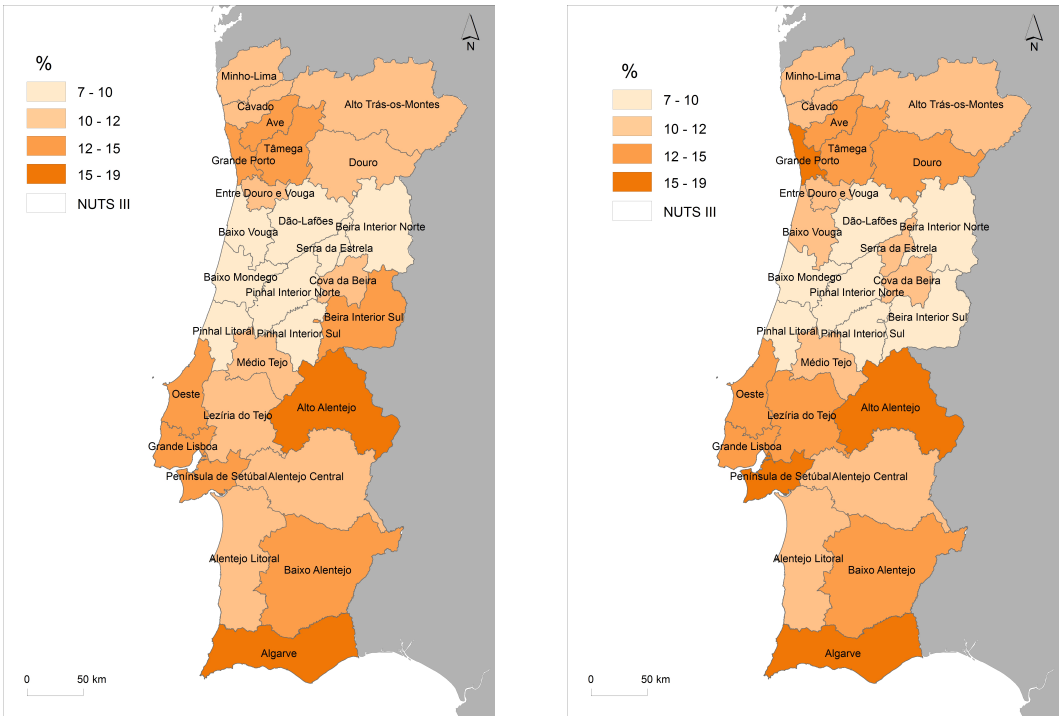

Figura 5.3: Taxa de desemprego (%) por NUTS III, relativa ao 1º trimestre de 2011 (à esquerda) e ao ano 2011 (à direita)

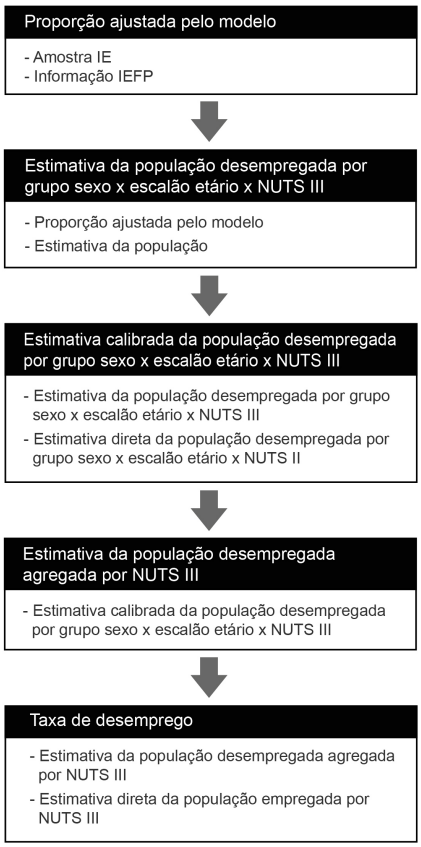

Figura 5.4: Esquema do Procedimento Prático

5.2 Indicadores de performance

5.2.1 Estimação da variância

A margem de erro das estimativas obtidas vai ser apresentada sob a forma de coeficiente de variação. Desta forma, é necessário efetuar o cálculo da variância.
Os estimadores que utilizámos são muito complexos e, como tal, a sua variância é muito difícil de expressar analiticamente. Vai usar-se a metodologia Bootstrap para a estimar.

Técnica de Bootstrap

A técnica de Bootstrap começou gradualmente a atrair a atenção de muitos estaticistas que trabalham com inquéritos amostrais, como alternativa para estimar a variância de estimadores complexos.
O método de Bootstrap é um método de reamostragem [6]. Seja n a dimensão da amostra original. Segundo este método, são retiradas B amostras de dimensão n, com reposição, da amostra original e calculam-se as estimativas $\hat{\theta}_b^*$, $b = 1, ..., B$ da mesma forma que $\hat{\theta}$ foi calculada.
A variância de $\hat{\theta}$ pode ser aproximada por:

$$var_B(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\bar{\theta}}^* \right)^2,$$

onde

$$\hat{\bar{\theta}}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*.$$

O estimador Bootstrap do coeficiente de variação (%) de $\hat{\theta}$ é:

$$cv_B(\hat{\theta}) = \frac{\sqrt{var_B(\hat{\theta})}}{\hat{\theta}} \times 100\%.$$

A reamostragem associada à técnica de Bootstrap foi aplicada separadamente a cada região NUTS III, de forma a garantir que cada subamostra tenha o mesmo número de elementos pertencentes a cada NUTS III, que corresponde ao número de elementos dessa região na amostra original.
A partir da amostra do IE do 1º trimestre de 2011, foram retiradas 1000 subamostras Bootstrap e obtiveram-se os valores aproximados dos coeficientes de variação das estimativas da população desempregada que estão apresentados na Tabela 5.4. Também foi feita uma reamostragem de Bootstrap a partir da amostra anual de 2011, sendo que os

NUTS II	NUTS III	CV (%) da população desempregada	
		Antes da calibração	Depois da calibração
Norte		4.1	0
	Minho-Lima	10.2	8.2
	Cávado	9.1	7.5
	Ave	7.9	6.4
	Grande Porto	6.1	4.3
	Tâmega	8.0	6.2
	Entre Douro e Vouga	10.2	8.5
	Douro	10.2	9.9
	Alto Trás-os-Montes	10.8	9.5
Centro		6.5	0
	Baixo Vouga	10.9	7.2
	Baixo Mondego	10.3	8.1
	Pinhal Litoral	10.8	7.7
	Pinhal Interior Norte	10.8	8.5
	Pinhal Interior Sul	10.8	8.9
	Dão-Lafões	10.2	6.9
	Serra da Estrela	9.8	7.4
	Beira Interior Norte	10.8	8.0
	Beira Interior Sul	9.8	7.5
	Cova da Beira	12.1	10.1
	Oeste	12.9	11.1
	Médio Tejo	12.2	8.4
Lisboa		4.8	0
	Grande Lisboa	5.4	2.5
	Península de Setúbal	8.2	5.8
Alentejo		6.3	0
	Alentejo Litoral	11.1	8.8
	Alto Alentejo	14.2	11.2
	Alentejo Central	10.0	7.5
	Baixo Alentejo	9.7	6.8
	Lezíria do Tejo	9.0	5.7
Algarve		5.3	0

Tabela 5.4: Coeficientes de variação (em %) das estimativas da população desempregada antes e depois da calibração, por NUTS III de Portugal Continental, referentes ao 1º trimestre de 2011

coeficientes de variação calculados para este período são apresentados na Tabela 5.5.

Note-se que, quando estimamos os coeficientes de variação de estimativas que foram calibradas por NUTS II, estes coeficientes são zero ao nível NUTS II. Afinal, ao obrigarmos a que as estimativas agregadas por NUTS II coincidam com as estimativas oficiais, estamos a exigir que as estimativas das várias réplicas Bootstrap sejam iguais ao nível NUTS II, o que implica que a estimativa da variância seja zero. Assim, não fará sentido comparar coeficientes de variação das estimativas model-based ao nível NUTS II com coeficientes de variação das estimativas oficiais a este nível.

Como era de esperar, a calibração provocou um decréscimo nos coeficientes de variação,

ao nível NUTS III. A razão está precisamente na condição que foi imposta às estimativas, reduzindo a variabilidade entre as réplicas Bootstrap. Note-se que as estimativas calibradas resultaram da multiplicação de uma constante pelas estimativas originais, em que essa constante é dada pelo quociente entre o total de desempregados oficial em determinada NUTS II e a soma do número de desempregados nas NUTS III que pertencem a essa NUTS II. No entanto, essa constante não é a mesma para todas as réplicas Bootstrap. Sabemos apenas que toma valores próximos de 1, o que evita, de facto, grande variabilidade.

		CV (%) da população desempregada	
NUTS II	**NUTS III**	**Antes da calibração**	**Depois da calibração**
Norte		**2.9**	**0**
	Minho-Lima	8.0	7.5
	Cávado	7.3	6.6
	Ave	6.4	5.9
	Grande Porto	4.5	3.4
	Tâmega	6.0	5.5
	Entre Douro e Vouga	8.1	7.7
	Douro	8.6	8.6
	Alto Trás-os-Montes	10.1	10.0
Centro		**4.4**	**0**
	Baixo Vouga	8.2	6.7
	Baixo Mondego	8.1	7.0
	Pinhal Litoral	7.9	7.4
	Pinhal Interior Norte	8.6	7.6
	Pinhal Interior Sul	7.7	6.4
	Dão-Lafões	8.2	7.3
	Serra da Estrela	7.6	7.3
	Beira Interior Norte	9.1	8.7
	Beira Interior Sul	7.9	7.3
	Cova da Beira	9.1	8.7
	Oeste	9.2	7.5
	Médio Tejo	11.1	8.9
Lisboa		**3.6**	**0**
	Grande Lisboa	4.1	2.0
	Península de Setúbal	5.9	4.8
Alentejo		**4.9**	**0**
	Alentejo Litoral	8.8	7.9
	Alto Alentejo	9.1	7.7
	Alentejo Central	8.0	7.0
	Baixo Alentejo	8.2	6.7
	Lezíria do Tejo	7.3	5.1
Algarve		**4.0**	**0**

Tabela 5.5: Coeficientes de variação (em %) das estimativas da população desempregada por NUTS III de Portugal Continental, referentes ao ano completo de 2011

Relativamente aos coeficientes de variação das estimativas ao nível NUTS III, podemos observar que as regiões Oeste e Alto Alentejo são as que apresentam maiores valores

(11.1% e 11.2%), no 1º trimestre de 2011. Estas duas regiões têm uma dimensão populacional pequena, facto que influencia naturalmente os valores dos coeficientes de variação, Figura 5.6. As regiões Grande Porto e Grande Lisboa são as que apresentam menores coeficientes de variação (4.3% e 2.5% respetivamente), e são as regiões com maior dimensão populacional. Estas quatro regiões destacam-se no gráfico referente ao 1º trimestre de 2011, Figura 5.5. No gráfico referente ao ano 2011, podemos ver que as regiões que apresentam menores coeficientes de variação são as mesmas, mas as regiões que apresentam maiores valores são Alto Trás-Os-Montes (10.0%) e Médio Tejo (8.9%), ambas com uma dimensão populacional pequena. Podemos ver melhor a relação entre os coeficientes de variação e a dimensão populacional das NUTS III, referentes ao ano 2011, através dos mapas apresentados na Figura 5.7.

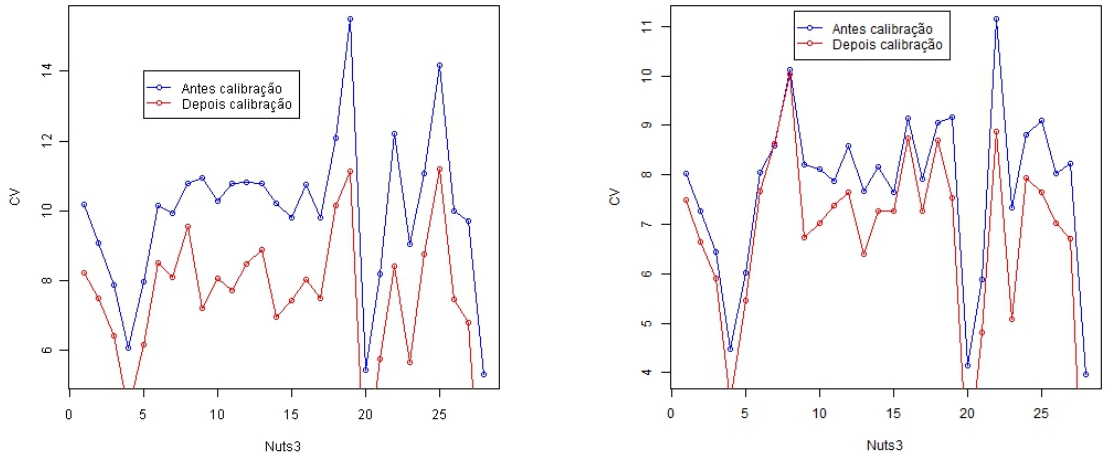

Figura 5.5: Estimativa do coeficiente de variação das estimativas obtidas antes e depois da calibração, por NUTS III, para o 1º trimestre de 2011 (à esquerda), e para o ano completo de 2011 (à direita)

Os boxplots representados na Figura 5.8 permitem ter uma ideia da distribuição destes valores. Note-se que há dois outliers no 1º trimestre de 2011, e quatro outliers no ano completo de 2011. Um dos outliers comuns corresponde ao Algarve e deve-se ao facto de esta região NUTS III coincidir com uma região NUTS II. Como vimos, os coeficientes de variação aproximados pelo método Bootstrap são zero para as regiões NUTS II. O outro outlier inferior comum corresponde à região Grande Lisboa, e deve-se provavelmente à sua grande dimensão populacional. O boxplot de 2011 mostra que há mais dois outliers: Grande Porto (outlier inferior) e Alto Trás-Os-Montes (outlier superior).

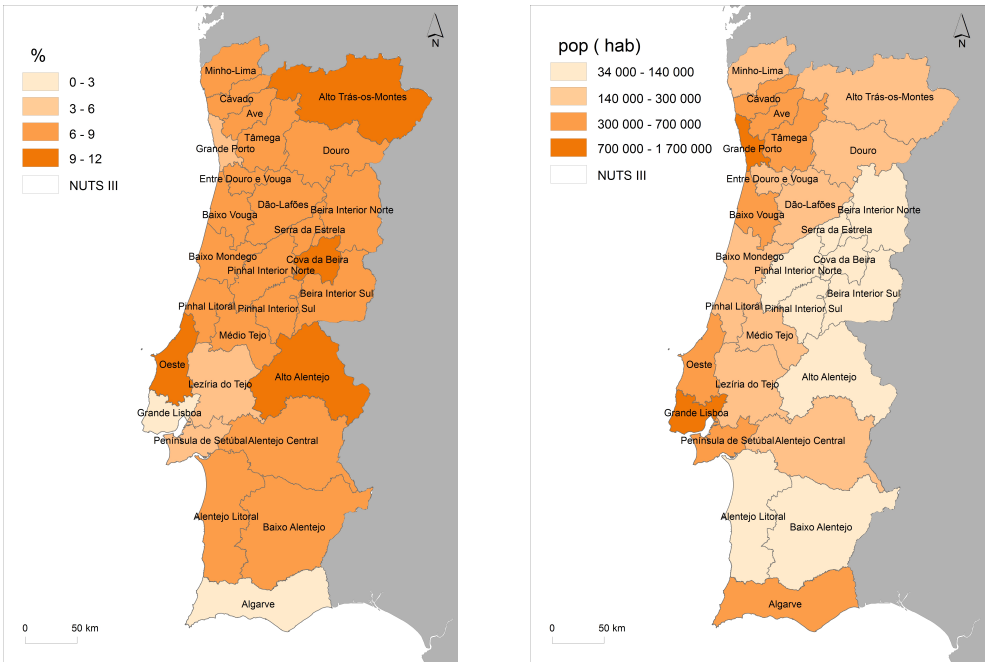

Figura 5.6: Coeficientes de variação (à esquerda) e estimativas da população com 15 ou mais anos (à direita) por NUTS III, referentes ao 1º trimestre de 2011

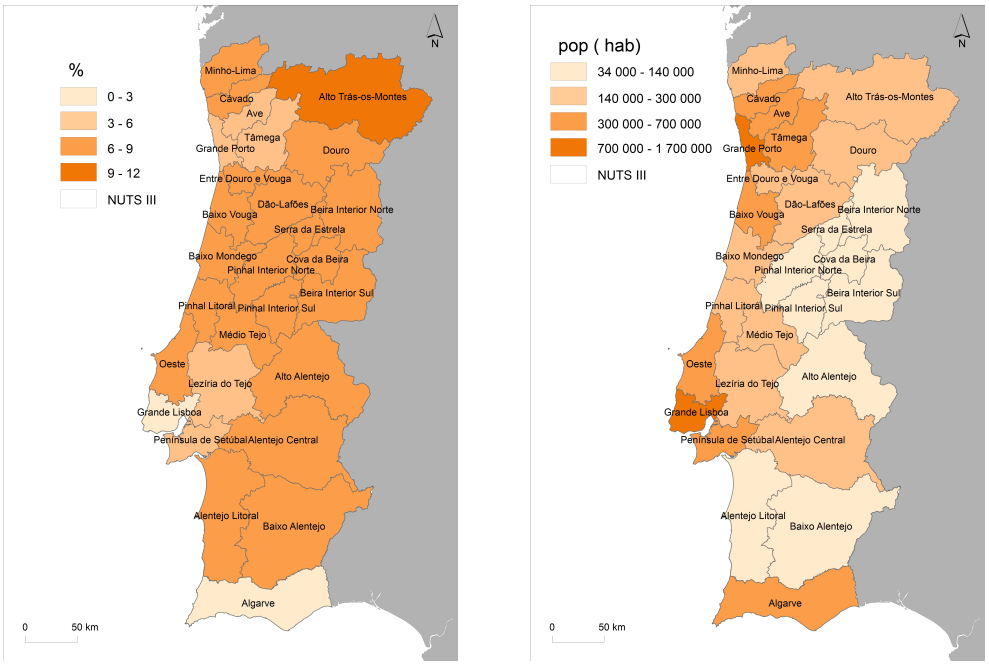

Figura 5.7: Coeficientes de variação (à esquerda) e estimativas da população com 15 ou mais anos (à direita) por NUTS III, referentes ao ano 2011

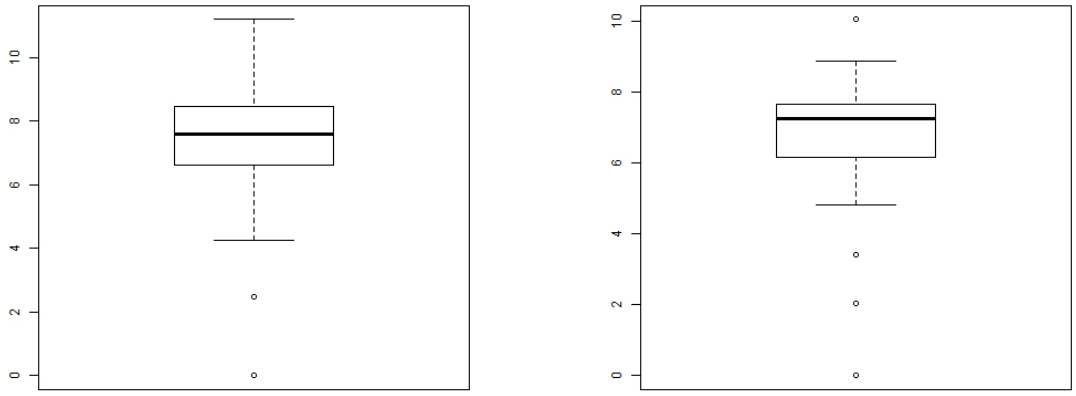

Figura 5.8: Boxplot dos coeficientes de variação das estimativas referentes ao 1º trimestre de 2011 (à esquerda), e referentes ao ano completo de 2011 (à direita)

Podemos comparar os coeficientes de variação das estimativas oficiais (produzidas pelo método direto) por NUTS II com os coeficientes de variação das estimativas model-based obtidas por NUTS III, tendo em conta a diferença do nível geográfico, Tabela 5.6. É claro que há uma forte tendência para que coeficientes de variação de estimativas obtidas para NUTS II sejam inferiores a coeficientes de variação de estimativas obtidas para NUTS III.

Nuts3	CV
Norte	4.3
Centro	6.7
Lisboa	5.4
Alentejo	8.1
Algarve	6.8

Tabela 5.6: Coeficientes de variação (em %) das estimativas da população desempregada calculadas pelo método direto, por NUTS III, para o 1º trimestre de 2011

Os coeficientes de variação das estimativas model-based obtidas para as regiões NUTS III são inferiores aos coeficientes das estimativas diretas, para todas as NUTS III, no 1º trimestre de 2011 (Figura 5.9). Já no ano completo de 2011, podemos notar que há uma região em que isso não acontece, Médio Tejo.

Pelo gráfico representado na Figura 5.10, podemos verificar que, para o 1º trimestre de 2011, os desvios-padrão das estimativas model-based são inferiores aos desvios-padrão das estimativas diretas, para todas as NUTS III. Já no ano completo de 2011, há uma região em que isso não se verifica, e provavelmente será o Médio Tejo também.

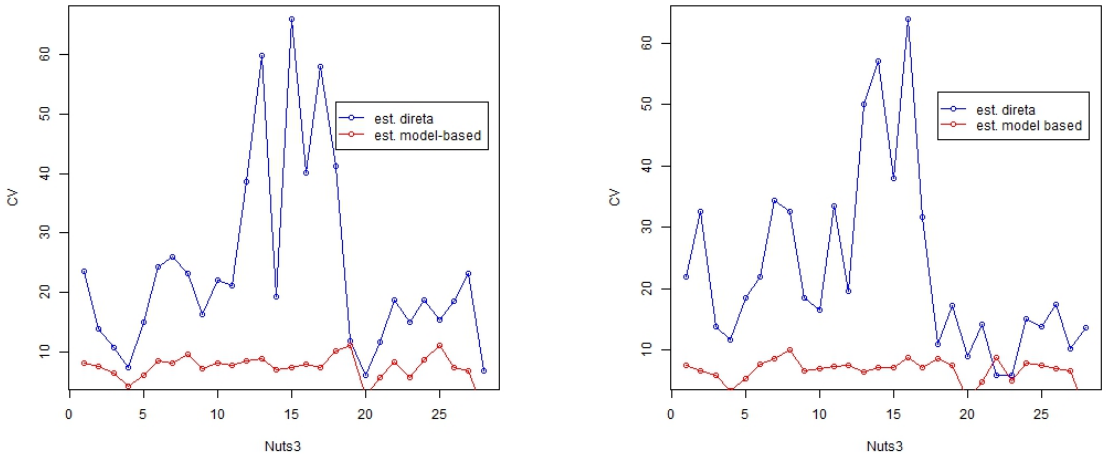

Figura 5.9: CV das estimativas diretas e CV das estimativas model-based, por NUTS III, referentes ao 1º trimestre de 2011 (à esquerda), e referentes a 2011 (à direita)

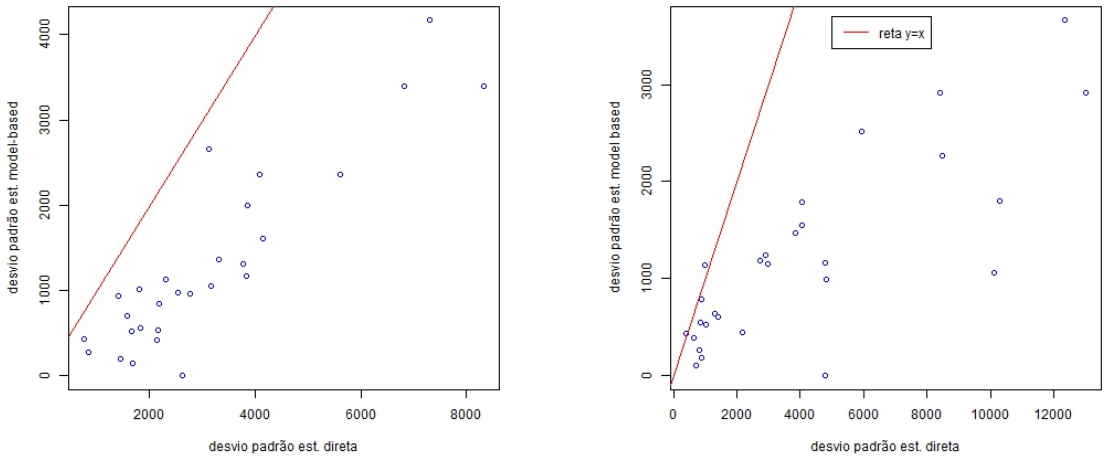

Figura 5.10: Desvio-padrão das estimativas diretas versus desvio-padrão das estimativas model-based, por NUTS III, referentes ao 1º trimestre de 2011 (à esquerda), e referentes a 2011 (à direita)

5.2.2 Outros indicadores

Além da variância e coeficiente de variação apresentados no capítulo anterior, serão calculados outros indicadores que também permitem avaliar as estimativas em questão. Estes também vão ser calculados a partir de métodos numéricos, com recurso a réplicas de Bootstrap.
Seja b ($b = 1, ..., B$) o índice da simulação, e $\hat{\theta}$ e $\hat{\theta}_b^*$ denotam respetivamente a estimativa do parâmetro de interesse calculada a partir da amostra original e a estimativa obtida na simulação b. Os critérios de avaliação usados neste estudo de simulação são:

- *Relative Bias* (RB) que pode ser aproximado por:

$$RB\% = \frac{1}{B}\left(\sum_{b=1}^{B}\frac{\hat{\theta}_b^* - \hat{\theta}}{\hat{\theta}}\right) \text{x}100\%$$

- *Relative Root Mean Squared Error* (RRMSE) que pode ser aproximado por:

$$RRMSE\% = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\frac{\hat{\theta}_b^* - \hat{\theta}}{\hat{\theta}}\right)^2} \text{x}100\%$$

Com base nos critérios de cima, ao nível NUTS III, expressos em percentagem, é possível considerar dois critérios de avaliação global:

- *Average Absolute Relative Bias* (AARB) , que corresponde ao cálculo da média dos valores absolutos do RB de todas as NUTS III d ($d = 1, ..., D$), em que D corresponde ao número total de NUTS III, e pode ser aproximado por:

$$AARB = \frac{1}{D}\sum_{d=1}^{D}|RB_d|$$

- *Average Relative Root Mean Squared Error* (ARRMSE), que corresponde ao cálculo da média dos valores de RRMSE de todas as NUTS III, e pode ser aproximado por:

$$ARRMSE = \frac{1}{D}\sum_{d=1}^{D}RRMSE_d$$

Aplicação dos indicadores

Nesta secção serão apresentados os resultados dos valores aproximados dos indicadores de performance mencionados.
Relativamente ao *Relative Bias* podemos notar, a partir do gráfico representado na Figura 5.11, que cerca de 50% dos valores deste indicador estão entre -5% e 1% , tanto para o 1º trimestre de 2011 como para o ano completo de 2011. A mediana é próxima de -1%. Os boxplots apresentados permitem concluir que as estimativas obtidas são aproximadamente centradas, pois o RB toma valores relativamente próximos de zero. No entanto, nota-se uma assimetria nas amostras correspondentes.

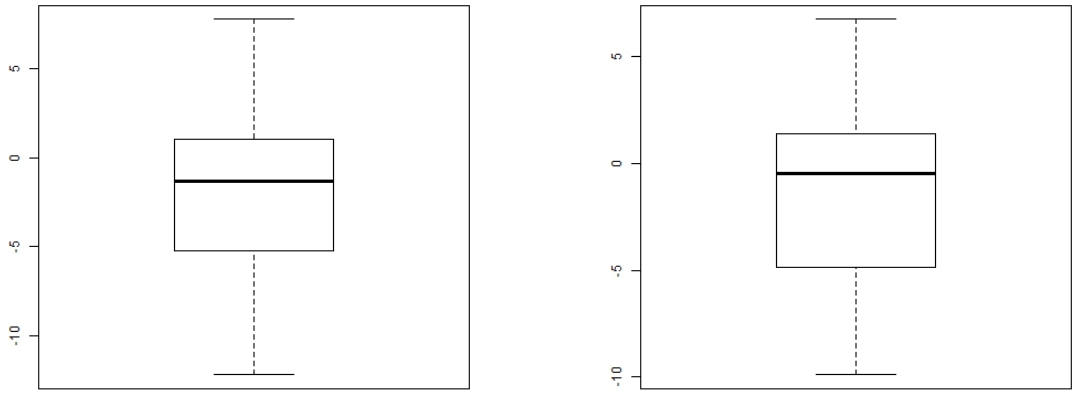

Figura 5.11: Boxplot das estimativas do *Relative Bias*(em %) referentes ao 1º trimestre de 2011 (à esquerda), e referentes ao ano completo de 2011 (à direita)

Pelos gráficos representados na Figura 5.12 podemos notar que há duas regiões com baixo RRMSE: Grande Porto (4.8%) e Grande Lisboa (2.5%). Mais uma vez, não estamos a incluir o Algarve. Destacam-se três regiões com maiores valores de RRMSE. No 1º trimestre de 2011 essas regiões correspondem a Pinhal Interior Sul (15.1%), Oeste (13.4%), e Alto Alentejo (13.6%). Para o ano de 2011, destacam-se Alto Trás-Os-Montes (14.1%), Beira Interior Norte (12.1%), e Médio Tejo (11.1%).
Este indicador é muito importante na avaliação da qualidade das estimativas, pois envolve variância e centricidade. O maior valor obtido é 15.1% para o 1º trimestre e 14.1% para o ano completo, e a média dos valores de RRMSE de todas as NUTS III é 8.5% para o 1º trimestre e é de 7.7% para 2011, o que parece bastante razoável.
Podemos consultar os valores destes dois indicadores na Tabela 5.7.

		RB (%)		RRMSE (%)	
NUTS II	**NUTS III**	**1T2011**	**2011**	**1T2011**	**2011**
Norte					
	Minho-Lima	-0.2	-0.7	8.2	7.5
	Cávado	1.3	-1.8	7.6	6.9
	Ave	1.2	2.6	6.5	6.4
	Grande Porto	2.3	1.2	4.8	3.6
	Tâmega	-2.0	-0.2	6.5	5.5
	Entre Douro e Vouga	1.6	5.5	7.8	9.9
	Douro	-3.3	-6.2	8.7	10.6
	Alto Trás-os-Montes	-8.0	-9.9	12.5	14.1
Centro					
	Baixo Vouga	0.6	2.1	7.2	7.0
	Baixo Mondego	-5.3	-4.1	9.6	8.1
	Pinhal Litoral	-1.9	0.1	8.0	7.4
	Pinhal Interior Norte	-5.9	-1.2	10.3	7.7
	Pinhal Interior Sul	-12.2	-5.2	15.1	8.2
	Dão-Lafões	4.2	-3.2	8.1	7.9
	Serra da Estrela	-4.7	-6.7	8.8	9.9
	Beira Interior Norte	-5.2	-8.3	9.6	12.1
	Beira Interior Sul	-6.1	-5.5	9.7	9.1
	Cova da Beira	-7.9	-5.7	12.8	10.4
	Oeste	7.5	4.4	13.4	8.7
	Médio Tejo	1.6	6.7	8.6	11.1
Lisboa					
	Grande Lisboa	-0.4	-0.1	2.5	2.0
	Península de Setúbal	0.9	0.2	5.8	4.8
Alentejo					
	Alentejo Litoral	-6.2	-4.3	10.7	9.0
	Alto Alentejo	7.8	4.5	13.6	8.9
	Alentejo Central	-3.1	-4.5	8.1	8.4
	Baixo Alentejo	-0.7	2.1	6.8	7.0
	Lezíria do Tejo	-0.3	0.7	5.7	5.1
Algarve					

Tabela 5.7: *Relative Bias*(em %) e *Relative Root Mean Squared Error*(em %) das estimativas da população desempregada por NUTS III de Portugal Continental, referentes ao 1º trimestre de 2011 e ao ano de 2011 (1º trimestre ao 4º trimestre)

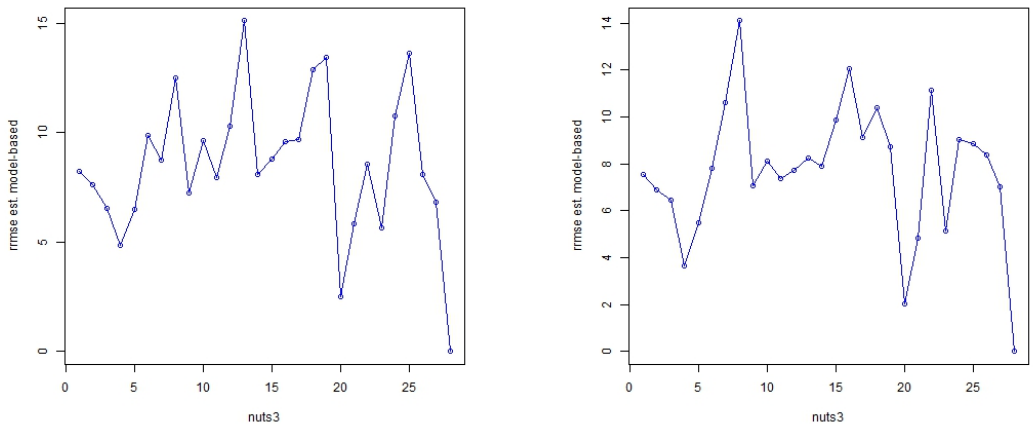

Figura 5.12: Estimativa do RRMSE por NUTS III, relativa ao 1º trimestre de 2011 (à esquerda), e relativa ao ano 2011 (à direita)

Capítulo 6

Discussão e Conclusões

Quando não é possível aumentar o tamanho da amostra de um pequeno domínio na fase de concepção, a qualidade das estimativas fica dependente da metodologia utilizada. Como a informação local não é suficiente para produzir estimativas diretas razoáveis, torna-se necessário "pedir informação emprestada" a outros domínios.

O modelo misto linear generalizado revelou ser uma boa solução para este problema. Foi usado um modelo ao nível área, uma vez que a informação auxiliar não estava disponível ao nível do indivíduo.

Através da aplicação da metodologia do ONS aos dados do IE referentes ao 1º trimestre de 2011 e ao ano completo de 2011, obtivemos estimativas para os totais e taxas de desemprego ao nível NUTS III, que são consistentes com as estimativas oficiais ao nível NUTS II. Os indicadores de performance revelam uma boa precisão e um bias relativamente pequeno.

As estimativas model-based produzidas distribuem-se aproximadamente sobre a reta $y = x$ relativamente às estimativas diretas (sendo estas centradas), o que permite concluir que as estimativas model-based são aproximadamente centradas.

A calibração feita às estimativas obtidas pelo modelo reduziu a variância e assegurou a consistência entre estas estimativas e as estimativas oficiais.

As estimativas obtidas pelo método indireto por NUTS III para o 1º trimestre de 2011 são mais precisas do que as estimativas obtidas pelo método direto, para todas as NUTS III. Relativamente ao ano completo 2011, acontece o mesmo, excepto para a região Médio Tejo.

Os coeficientes de variação obtidos para as estimativas model-based não excedem 11.2% (Algarve) para o 1º trimestre de 2011 e 8.9% (Médio Tejo) para o ano 2011, o que parece bom, e os menores valores correspondem às regiões Grande Porto e Grande Lisboa. Estas duas regiões têm uma maior dimensão populacional, daí apresentarem estimativas com

maior precisão.

Há muitos outros métodos que estão a ser estudados com a mesma finalidade. Os resultados dos indicadores de performance são muito importantes na avaliação da qualidade das estimativas, no entanto, a escolha do melhor método não é assim tão linear. Como vimos, é necessário ter em conta a relevância dos pressupostos iniciais, a disponibilidade e confiabilidade da informação auxiliar, facilidade de implementação, estabilidade das conclusões no tempo (fazendo o estudo para vários trimestres e anos), entre outros factores. A metodologia usada poderá vir a ser o "caminho a seguir" para, num futuro próximo, serem publicadas estimativas trimestrais e/ou anuais para a taxa de desemprego ao nível NUTS III.

Como trabalho futuro, seria interessante introduzir alguma componente espacial no modelo, uma vez que no desemprego o agrupamento geográfico tem com certeza influência. Desta forma, a "força emprestada" por outros domínios seria maior nas regiões vizinhas, tal como fará sentido.

Bibliografia

[1] Rao, J. N. K., *Small Area Estimation*, Wiley, 2003.

[2] Sarndal, C., Swensson, B., Wretman, J., *Model Assisted Survey Sampling*, Springer-Verlag, 1992.

[3] Kleinbaum, D. G., Klein, M., *Logistic Regression*, Springer, 2002.

[4] Hastings, D., Main, N., Brown, G. and Cruddas, M., *Development of improved estimation methods for local area unemployment levels and rates*, Labour Market Division, 2003.

[5] Longford, N. T., *Missing Data and Small-Area Estimation*, Springer, 2005.

[6] Saralegui, J., Herrador, M., Morales, D., Pérez, A., *Small Area Estimation in the Spanish Labour Force Survey*, Statistics in Transition, Vol. 7, No. 3, pp.571-585, 2005.

[7] Rao, J. N. K., *Inferential Issues in Small Area Estimation: Some New Developments*, Statistics in Transition, Vol. 7, No. 3, pp. 513-526, 2005.

[8] Office for National Statistics, *Model-Based Estimates of ILO Unemployment for LAD/UAs in Great Britain*, Guide for Users, 2006.

[9] *Introdution to Small Area Estimation*, SAE package developers, 2007.

[10] Srivastava, A. K., Sud, U. C., Chandra, K., *Small Area Estimation - An Application to National Sample Survey Data*, J. Ind. Soc. Agril. Statist., 61(2), 249-254, 2007.

[11] Pferffermann, D., Rao, C. R., *Sample Surveys: Inference and Analysis*, 2009.

[12] *Model-Based Estimates of ILO Unemployment for Parliamentary Constituencies in Great Britain Guide for Users*, 2009.

[13] Curtis, D., *Area random effects model-based estimation of unemployment at local authority level*, 7th Meeting of the National Methodology Advisory Committee.

[14] Correia, L., *Estimação em Pequenos Domínios, Aplicação do Estimador Sintético de Regressão ao Inquérito ao Emprego*, Instituto Nacional de Estatística, 2010.

[15] D'Alo, M., Consiglio, L., Falorsi, S., Ranalli, M. G., Solari, F., *Use of Spatial Information in Small Area Models for Unemployment Rate Estimation at Sub-Provincial Areas in Italy*, Journal of the Indian Society of Agrcultural Statistics, 43-53, 2012.

Anexos

NUTS

	Região NUTS III		
	Código	Designação	Região NUTS II
1	111	Minho-Lima	Norte
2	112	Cávado	
3	113	Ave	
4	114	Grande Porto	
5	115	Tâmega	
6	116	Entre Douro e Vouga	
7	117	Douro	
8	118	Alto Trás-os-Montes	
9	161	Baixo Vouga	Centro
10	162	Baixo Mondego	
11	163	Pinhal Litoral	
12	164	Pinhal Interior Norte	
13	166	Pinhal Interior Sul	
14	165	Dão-Lafões	
15	167	Serra da Estrela	
16	168	Beira Interior Norte	
17	169	Beira Interior Sul	
18	16A	Cova da Beira	
19	16B	Oeste	
20	171	Grande Lisboa	Lisboa
21	172	Península de Setúbal	
22	16C	Médio Tejo	Centro
23	185	Lezíria do Tejo	Alentejo
24	181	Alentejo Litoral	
25	182	Alto Alentejo	
26	183	Alentejo Central	
27	184	Baixo Alentejo	
28	150	Algarve	Algarve

Tabela 6.1: NUTS II e NUTS III de Portugal Continental

Software desenvolvido em R

Estimativas trimestrais

Total de desempregados na amostra por grupo sexo x escalão etário e NUTS III

```
> totais <- read.table("E:\\1t11\\dados\\totais_desemp.txt")
> totais_y <- c(totais[, 1], totais[, 2], totais[, 3], totais[,
+     4], totais[, 5], totais[, 6], totais[, 7], totais[, 8])
```

Número de inquiridos por grupo sexo x escalão etário e NUTS III

```
> inquiridos <- read.table("E:\\1t11\\dados\\ninquiridos.txt")
> ninquiridos <- c(inquiridos[, 1], inquiridos[, 2], inquiridos[,
+     3], inquiridos[, 4], inquiridos[, 5], inquiridos[, 6], inquiridos[,
+     7], inquiridos[, 8])
```

Total da população por grupo sexo x escalão etário e NUTS III

```
> pop <- read.table("E:\\1t11\\dados\\totais_pop.txt")
> n <- c(pop[, 1], pop[, 2], pop[, 3], pop[, 4], pop[, 5], pop[,
+     6], pop[, 7], pop[, 8])
```

Proporção de desempregados inscritos nos centros de emprego por grupo sexo x escalão etário e NUTS III

```
> iefp_grupo <- read.table("E:\\1t11\\dados\\iefp_grupo.txt")
> iefp <- c(iefp_grupo[, 1], iefp_grupo[, 2], iefp_grupo[, 3],
+     iefp_grupo[, 4], iefp_grupo[, 5], iefp_grupo[, 6], iefp_grupo[,
+         7], iefp_grupo[, 8])
> logit_iefp <- log(iefp/(1 - iefp))
```

Variáveis dummy que entram no modelo

- Indicador do grupo sexo x escalão etário

  ```
  > variavel <- c(rep(1, 28), rep(2, 28), rep(3, 28), rep(4, 28),
  +     rep(5, 28), rep(6, 28), rep(7, 28), rep(8, 28))
  > niveis <- 1:8
  > dummy_grupo <- factor(variavel, levels = niveis)
  ```

- Indicador da NUTS II

  ```
  > var2 <- rep(c(rep(1, 8), rep(2, 11), rep(3, 2), 2, rep(4, 5),
  +     rep(5, 1)), 8)
  > niveis2 <- 1:5
  > dummy_nuts2 <- factor(var2, levels = niveis2)
  ```

Dados que entram no modelo

```
> nuts3 <- rep(1:28, 8)
> dados <- data.frame(dummy_nuts2, nuts3, totais_y, ninquiridos,
+     dummy_grupo, logit_iefp)
```

Aplicação de um modelo logit com efeitos aleatórios

```
> library(lme4)
> modelo <- glmer(cbind(totais_y, ninquiridos - totais_y) ~ dummy_grupo +
+     logit_iefp + dummy_nuts2 + (1 | nuts3), data = dados, family = binomial("logit"))
> summary(modelo)


Generalized linear mixed model fit by the Laplace approximation
Formula: cbind(totais_y, ninquiridos - totais_y) ~ dummy_grupo + logit_iefp +      dummy_nuts2 + (1 | nuts3)
   Data: dados
   AIC   BIC logLik deviance
 277.5 325.2 -124.7    249.5
Random effects:
 Groups Name        Variance Std.Dev.
 nuts3  (Intercept) 0.010543 0.10268
Number of obs: 224, groups: nuts3, 28

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.64582    0.30298  -5.432 5.57e-08 ***
dummy_grupo2 -0.18315    0.11049  -1.658   0.0974 .
dummy_grupo3  1.07668    0.10355  10.397  < 2e-16 ***
dummy_grupo4  0.95993    0.11194   8.575  < 2e-16 ***
dummy_grupo5  0.70588    0.09420   7.494 6.70e-14 ***
dummy_grupo6  0.57979    0.09792   5.921 3.20e-09 ***
dummy_grupo7 -0.02032    0.10466  -0.194   0.8460
dummy_grupo8 -0.28567    0.12259  -2.330   0.0198 *
logit_iefp    0.52778    0.11490   4.593 4.36e-06 ***
dummy_nuts22 -0.06633    0.10040  -0.661   0.5088
dummy_nuts23  0.20367    0.11006   1.851   0.0642 .
dummy_nuts24  0.09389    0.09958   0.943   0.3458
dummy_nuts25  0.24156    0.13258   1.822   0.0685 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) dmmy_2 dmmy_3 dmmy_4 dmmy_5 dmmy_6 dmmy_7 dmmy_8 lgt_fp
dummy_grup2 -0.452
dummy_grup3 -0.248  0.488
dummy_grup4 -0.562  0.558  0.492
dummy_grup5 -0.298  0.544  0.552  0.552
dummy_grup6 -0.523  0.599  0.552  0.635  0.616
dummy_grup7  0.228  0.331  0.455  0.275  0.489  0.368
dummy_grup8  0.420  0.210  0.369  0.134  0.390  0.230  0.659
logit_iefp   0.954 -0.306 -0.082 -0.426 -0.117 -0.358  0.415  0.591
dummy_nts22  0.292 -0.127 -0.039 -0.190 -0.046 -0.146  0.173  0.250  0.422
dummy_nts23  0.104 -0.057 -0.020 -0.092 -0.018 -0.065  0.095  0.137  0.217
dummy_nts24  0.111 -0.067 -0.023 -0.098 -0.023 -0.072  0.097  0.141  0.233
dummy_nts25 -0.254  0.061  0.028  0.078  0.029  0.077 -0.059 -0.087 -0.175
            dmm_22 dmm_23 dmm_24
dummy_grup2
dummy_grup3
dummy_grup4
dummy_grup5
dummy_grup6
dummy_grup7
```

```
dummy_grup8
logit_iefp
dummy_nts22
dummy_nts23  0.395
dummy_nts24  0.434  0.357
dummy_nts25  0.178  0.192  0.213
```

Proporção ajustada de desempregados por grupo sexo x escalão etário e NUTS III

```
> prop_ajustada <- fitted(modelo)
```

Estimação do total de desempregados por grupo sexo x escalão etário e NUTS III

```
> total_desemp_grupo <- n * prop_ajustada
```

Calibração por grupo sexo x escalão etário e NUTS III, a partir das estimativas diretas

```
> estimativas_diretas <- read.table("E:\\1t11\\dados\\estimativas_diretas_nuts2_grupo_1t11.txt")
> estimativas_diretas <- estimativas_diretas[, 1]
> pe <- 0
> i = 1
> while (i <= 36) {
+     vetor <- c(rep(estimativas_diretas[i], 8), rep(estimativas_diretas[i +
+         1], 11), rep(estimativas_diretas[i + 2], 2), estimativas_diretas[i +
+         1], rep(estimativas_diretas[i + 3], 5), estimativas_diretas[i +
+         4])
+     pe <- append(pe, vetor)
+     i <- i + 5
+ }
> pe <- pe[2:225]
> pe2 <- 0
> j = 1
> while (j <= 197) {
+     vetor2 <- c(rep(sum(total_desemp_grupo[j:(j + 7)]), 8), rep(sum(total_desemp_grupo[(j +
+         8):(j + 18)]) + total_desemp_grupo[j + 21], 11), rep(sum(total_desemp_grupo[(j +
+         19):(j + 20)]), 2), sum(total_desemp_grupo[(j + 8):(j +
+         18)]) + total_desemp_grupo[j + 21], rep(sum(total_desemp_grupo[(j +
+         22):(j + 26)]), 5), rep(total_desemp_grupo[j + 27], 1))
+     pe2 <- append(pe2, vetor2)
+     j <- j + 28
+ }
> pe2 <- pe2[2:225]
> lambda2 <- pe/pe2
> consistentes_grupo <- total_desemp_grupo * lambda2
```

Confirmação da consistência das estimativas com as estimativas diretas ao nível dos grupos sexo x escalão etário e NUTS II

```
> c(sum(consistentes_grupo[1:8]), estimativas_diretas[1])
```

```
[1] 21479.26 21479.26
```

```
> c(sum(consistentes_grupo[9:19]) + consistentes_grupo[22], estimativas_diretas[2])
```

```
[1] 9847.899 9847.899
```

```
> c(sum(consistentes_grupo[20:21]), estimativas_diretas[3])
```

```
[1] 17559.58 17559.58
```

```
> c(sum(consistentes_grupo[23:27]), estimativas_diretas[4])
```

```
[1] 4522.737 4522.737
```

```
> c(consistentes_grupo[28], estimativas_diretas[5])
```

```
[1] 4237.242 4237.242
```

Agregação das estimativas por NUTS III

```
> total_nuts3_g <- 0
> for (j in 1:28) {
+     total_nuts3_g[j] <- consistentes_grupo[j]
+     i = j + 28
+     while (i <= 224) {
+         total_nuts3_g[j] <- total_nuts3_g[j] + consistentes_grupo[i]
+         i <- i + 28
+     }
+ }
> total_nuts3_g
```

```
 [1]  14174.041  26623.405  36675.856  97725.622  38436.812  15992.051
 [7]  13007.762  11896.232  22264.981  16251.952  12399.007   6315.124
[13]   1593.882  13418.282   2579.889   5154.031   3669.937   5142.750
[19]  23836.792 136468.216  58923.961  11544.524  14892.298   4972.108
[25]   9006.546   9388.380   8265.536  38628.741
```

Gráfico que relaciona as estimativas diretas com as estimativas model-based

```
est_direta_nuts3<-read.table("E:\\1t11\\dados\\estimativas_diretas_nuts3.txt")
est_direta_nuts3<-est_direta_nuts3[,1]
est_model_based<-total_nuts3_g

jpeg(filename = "est_direct_model.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(est_direta_nuts3,est_model_based,col="blue",xlab="estimativa direta",ylab="estimativa model-based")
abline(a=0,b=1,col="red")
legend(20000,80000, pt.bg="white", lty=1,"reta y=x",col="red")
dev.off()
```

Cálculo da taxa de desemprego por NUTS III

```
estimativa_empregados<-read.table("E:\\1t11\\dados\\est_empreg_nuts3.txt")
estimativa_empregados

estimativa_empreg<-estimativa_empregados[,1]

taxa_desemprego<-total_nuts3_g/(total_nuts3_g+estimativa_empreg)
taxa_desemprego
taxa_desemprego<-100*taxa_desemprego
```

Gráfico da taxa de desemprego por NUTS III

```
xx<-1:28
jpeg(filename = "taxa_desemp.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,taxa_desemprego,type="o",col="blue",xlab="nuts3",ylab="taxa de desemprego")
dev.off()
```

Agregação das estimativas por NUTS II

```
> total_nuts2_g <- 0
> for (i in 1:8) {
+     total_nuts2_g[1] <- total_nuts2_g[1] + total_nuts3_g[i]
+ }
> total_nuts2_g[2] <- 0
> for (i in 9:19) {
+     total_nuts2_g[2] <- total_nuts2_g[2] + total_nuts3_g[i]
+ }
> total_nuts2_g[2] <- total_nuts2_g[2] + total_nuts3_g[22]
> total_nuts2_g[3] <- 0
> for (i in 20:21) {
+     total_nuts2_g[3] <- total_nuts2_g[3] + total_nuts3_g[i]
+ }
> total_nuts2_g[4] <- 0
> for (i in 23:27) {
+     total_nuts2_g[4] <- total_nuts2_g[4] + total_nuts3_g[i]
+ }
> total_nuts2_g[5] <- 0
> total_nuts2_g[5] <- total_nuts3_g[28]
> total_nuts2_g

[1] 254531.78 124171.15 195392.18  46524.87  38628.74
```

Confirmação da consistência das estimativas obtidas por NUTS III com as estimativas diretas obtidas por NUTS II

```
> total_nuts2_g

[1] 254531.78 124171.15 195392.18  46524.87  38628.74

> estimativas_diretas_nuts2 <- read.table("E:\\1t11\\dados\\estimativas_diretas_desemp.txt")
> estimativas_diretas_nuts2[, 1]

[1] 254531.78 124171.15 195392.18  46524.87  38628.74
```

```
> estimativas_diretas_nuts2[, 1]

[1] 254531.78 124171.15 195392.18  46524.87  38628.74
```

Aplicação do método de Bootstrap

```
dados<-read.table("E:\\1t11\\dados\\dados.txt",header=T)
dados[1:5,]

linhas<-length(dados$CPT)
codigo<-1
for (i in 2:linhas){
if (dados$NUTS3[i]==dados$NUTS3[i-1])
{codigo[i]<-codigo[i-1]}
else {codigo[i]<-codigo[i-1]+1}
}

dados$cod<-codigo
dados[2000:2014, ]

leng<-0
pe2<-0
pe<-0
total_nuts3_g2<-0
est<-0
B<-1000

for (j in 1:B){
ordens_final<-0
pin<-1
for (i in 1:28){
leng[i]<-length(dados$CPT[which(dados$cod==i)])
b<-pin:(pin+leng[i]-1)
pin<-pin+leng[i]
ordens<-sample(b,leng[i],replace=TRUE)
ordens_final<-append(ordens_final, ordens)
}
ordens_final<-ordens_final[2:length(ordens_final)]
dados_desemp<-dados[c(ordens_final),]
tabela<-tapply(dados_desemp$CPT,dados_desemp$SEXO,length)
tabela2<-tapply(dados_desemp$CPT,dados_desemp$escalao,length)
tabela33<-tapply(dados_desemp$CPT,list(dados_desemp$see,dados_desemp$NUTS3),length)
dados_finais_<-data.frame(tabela33)
dados_f_<-replace(dados_finais_, is.na(dados_finais_), 0)
dados_f2_<-as.matrix(dados_f_)
dados_f3_<-t(dados_f2_)
ordem_<-c(dados_f3_[1:8,2],dados_f3_[10:20,2],dados_f3_[22:23,2],
dados_f3_[21,2],dados_f3_[28,2],dados_f3_[24:27,2],dados_f3_[9,2],
dados_f3_[1:8,3],dados_f3_[10:20,3],dados_f3_[22:23,3],
dados_f3_[21,3],dados_f3_[28,3],dados_f3_[24:27,3],dados_f3_[9,3],
dados_f3_[1:8,4],dados_f3_[10:20,4],dados_f3_[22:23,4],
dados_f3_[21,4],dados_f3_[28,4],dados_f3_[24:27,4],dados_f3_[9,4],
dados_f3_[1:8,5],dados_f3_[10:20,5],dados_f3_[22:23,5],
dados_f3_[21,5],dados_f3_[28,5],dados_f3_[24:27,5],dados_f3_[9,5],
dados_f3_[1:8,6],dados_f3_[10:20,6],dados_f3_[22:23,6],
```

```
dados_f3_[21,6],dados_f3_[28,6],dados_f3_[24:27,6],dados_f3_[9,6],
dados_f3_[1:8,7],dados_f3_[10:20,7],dados_f3_[22:23,7],
dados_f3_[21,7],dados_f3_[28,7],dados_f3_[24:27,7],dados_f3_[9,7],
dados_f3_[1:8,8],dados_f3_[10:20,8],dados_f3_[22:23,8],
dados_f3_[21,8],dados_f3_[28,8],dados_f3_[24:27,8],dados_f3_[9,8],
dados_f3_[1:8,9],dados_f3_[10:20,9],dados_f3_[22:23,9],
dados_f3_[21,9],dados_f3_[28,9],dados_f3_[24:27,9],dados_f3_[9,9]
)

dados_f4_<-as.vector(ordem_)
ninquiridos2<-dados_f4_

dados2<-dados_desemp[dados_desemp$CPT==2,]

tabela3<-tapply(dados2$CPT,list(dados2$see,dados2$NUTS3),length)

dados_finais<-data.frame(tabela3)

dados_f<-replace(dados_finais, is.na(dados_finais), 0)

dados_f2<-as.matrix(dados_f)

dados_f3<-t(dados_f2)

ordem<-c(dados_f3[1:8,1],dados_f3[10:20,1],dados_f3[22:23,1],
dados_f3[21,1],dados_f3[28,1],dados_f3[24:27,1],dados_f3[9,1],
dados_f3[1:8,2],dados_f3[10:20,2],dados_f3[22:23,2],
dados_f3[21,2],dados_f3[28,2],dados_f3[24:27,2],dados_f3[9,2],
dados_f3[1:8,3],dados_f3[10:20,3],dados_f3[22:23,3],
dados_f3[21,3],dados_f3[28,3],dados_f3[24:27,3],dados_f3[9,3],
dados_f3[1:8,4],dados_f3[10:20,4],dados_f3[22:23,4],
dados_f3[21,4],dados_f3[28,4],dados_f3[24:27,4],dados_f3[9,4],
dados_f3[1:8,5],dados_f3[10:20,5],dados_f3[22:23,5],
dados_f3[21,5],dados_f3[28,5],dados_f3[24:27,5],dados_f3[9,5],
dados_f3[1:8,6],dados_f3[10:20,6],dados_f3[22:23,6],
dados_f3[21,6],dados_f3[28,6],dados_f3[24:27,6],dados_f3[9,6],
dados_f3[1:8,7],dados_f3[10:20,7],dados_f3[22:23,7],
dados_f3[21,7],dados_f3[28,7],dados_f3[24:27,7],dados_f3[9,7],
dados_f3[1:8,8],dados_f3[10:20,8],dados_f3[22:23,8],
dados_f3[21,8],dados_f3[28,8],dados_f3[24:27,8],dados_f3[9,8]
)

dados_f4<-as.vector(ordem)

ff<-data.frame(dummy_nuts2,nuts3,dados_f4,ninquiridos2,dummy_grupo,logit_iefp)

modelo<-glmer(cbind(dados_f4,ninquiridos2-dados_f4)~dummy_grupo+logit_iefp+dummy_nuts2+(1|nuts3),dados=ff,family=binomial("log

ajustados<-fitted(modelo)

total_desemp_grupo<-n*ajustados

i=1
while(i<=36){
vetor<-c(
rep(estimativas_diretas[i],8),
```

```
rep(estimativas_diretas[i+1],11),
rep(estimativas_diretas[i+2],2),
estimativas_diretas[i+1],
rep(estimativas_diretas[i+3],5),
estimativas_diretas[i+4])
pe<-append(pe,vetor)
i<-i+5}

pe<-pe[2:225]
j=1
while (j<=197){
vetor2<-c(
rep(sum(total_desemp_grupo[j:(j+7)]),8),
rep(sum(total_desemp_grupo[(j+8):(j+18)])+total_desemp_grupo[j+21],11),
rep(sum(total_desemp_grupo[(j+19):(j+20)]),2),
sum(total_desemp_grupo[(j+8):(j+18)])+total_desemp_grupo[j+21],
rep(sum(total_desemp_grupo[(j+22):(j+26)]),5),
rep(total_desemp_grupo[j+27],1))
pe2<-append(pe2,vetor2)
j<-j+28}

pe2<-pe2[2:225]

lambda2<-pe/pe2

consistentes_grupo<-total_desemp_grupo*lambda2

for (j in 1:28){
total_nuts3_g2[j]<-consistentes_grupo[j]
i=j+28
while (i<=224){
total_nuts3_g2[j]<-total_nuts3_g2[j]+consistentes_grupo[i]
i<-i+28}}

est<-append(est,total_nuts3_g2)

}

limite<-28*B+1
est<-est[2:limite]
matriz<-matrix(est,nrow=28,ncol=B)
```

Cálculo da variância, coeficiente de variação e outros indicadores de performance

```
media_boot<-0
var_boot<-0
rb<-0
rrmse<-0
arrmse<-0
for (i in 1:28){
media_boot[i]<-mean(matriz[i,])
var_boot[i]<-sum((matriz[i,]-media_boot[i])^2)/(B-1)
rb[i]<-sum((matriz[i,]-total_nuts3_g[i])/total_nuts3_g[i])*100/B
rrmse[i]<-100*sqrt((1/B)*sum(((matriz[i,]-total_nuts3_g[i])/total_nuts3_g[i])^2))
```

```
}

arrmse<-(1/28)*sum(rrmse)
var_boot
rb
rrmse
arrmse

cv<-(sqrt(var_boot)/total_nuts3_g)*100
cv
```

Alguns gráficos de indicadores de performance

```
jpeg(filename = "boxplot_cv.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
boxplot(cv)
dev.off()

se_boot<-sqrt(var_boot)
se_boot

jpeg(filename = "se.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,se_boot,type="o",,col="blue",xlab="nuts3",ylab="desvio padrão est. model-based")
dev.off()

jpeg(filename = "cv.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,cv,type="o",,col="blue",xlab="nuts3",ylab="cv est. model-based")
dev.off()

se_direto<-read.table("E:\\1t11\\dados\\se_desemp_direto.txt",header=F)
se_direto<-se_direto[,1]

cv_direto<-read.table("E:\\1t11\\dados\\cv_direto.txt",header=F)
cv_direto<-cv_direto[,1]

jpeg(filename = "se_direct_model.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(se_direto,se_boot,xlab="desvio padrão est. direta", ylab="desvio padrão est. model-based",col="blue")
abline(a=0,b=1,col="red")
legend(2000,5000, pt.bg="white", lty=1,"reta y=x",col="red")
dev.off()

jpeg(filename = "cv_direct_model.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,cv_direto,type="o",col="blue",xlab="Nuts3",ylab="CV")
lines(cv,type="o",pch=21, lty=1,col="red")
legend(18,52,pch=21, pt.bg="white", lty=1,c("est. direta","est. model-based"),col=c("blue","red"))
```

```
dev.off()

jpeg(filename = "rrmse.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,rrmse,type="o",,col="blue",xlab="nuts3",ylab="rrmse est. model-based")
dev.off()

jpeg(filename = "boxplot_rb.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
boxplot(rb)
dev.off()
```

Estimativas Anuais

Amostra anual sem a rotação 5

```
> dados_1t<-read.table(paste("D:\\problema\\2011\\IE_pais_1t11.csv",sep=""),head=T,sep=",",dec=".")
> dados_1t_anual_1<-dados_1t[which(dados_1t$ROTACAO==6409),]
> dados_1t_anual_2<-dados_1t[which(dados_1t$ROTACAO==1114),]
> dados_1t_anual<-rbind(dados_1t_anual_1,dados_1t_anual_2)
> dados_2t<-read.table(paste("D:\\problema\\2011\\IE_pais_2t11.csv",sep=""),head=T,sep=",",dec=".")
> dados_2t_anual_1<-dados_2t[which(dados_2t$ROTACAO==6110),]
> dados_2t_anual_2<-dados_2t[which(dados_2t$ROTACAO==2115),]
> dados_2t_anual<-rbind(dados_2t_anual_1,dados_2t_anual_2)
> dados_3t<-read.table(paste("D:\\problema\\2011\\IE_pais_3t11.csv",sep=""),head=T,sep=",",dec=".")
> dados_3t_anual_1<-dados_3t[which(dados_3t$ROTACAO==6210),]
> dados_3t_anual_2<-dados_3t[which(dados_3t$ROTACAO==3116),]
> dados_3t_anual<-rbind(dados_3t_anual_1,dados_3t_anual_2)
> dados_4t<-read.table(paste("D:\\problema\\2011\\IE_pais_4t11.csv",sep=""),head=T,sep=",",dec=".")
> dados_4t<-dados_4t[,1:64]
> dados_4t_anual_1<-dados_4t[which(dados_4t$ROTACAO==6310),]
> dados_4t_anual_2<-dados_4t[which(dados_4t$ROTACAO==4111),]
> dados_4t_anual<-rbind(dados_4t_anual_1,dados_4t_anual_2)
> dados_anual_semrot5<-rbind(dados_1t_anual,dados_2t_anual,dados_3t_anual,dados_4t_anual)
```

Rotação 5

```
> l1<-length(dados_1t$ESTRATO)
> l2<-length(dados_2t$ESTRATO)
> l3<-length(dados_3t$ESTRATO)
> l4<-length(dados_4t$ESTRATO)
> dados_1t$trimestre<-rep(1,l1)
> dados_2t$trimestre<-rep(1,l2)
> dados_3t$trimestre<-rep(1,l3)
> dados_4t$trimestre<-rep(1,l4)
> total_dados<-rbind(dados_1t,dados_2t,dados_3t,dados_4t)
> sort.dados<-total_dados[order(total_dados$AREA_AM_ORIG), ]
> linhas<-length(sort.dados$AREA_AM_ORIG)
> codigo<-1
> for (i in 2:linhas){
+ if (sort.dados$AREA_AM_ORIG[i]==sort.dados$AREA_AM_ORIG[i-1])
+ {codigo[i]<-codigo[i-1]}
```

```
+ else {codigo[i]<-codigo[i-1]+1}
+ }
> l<-length(unique(codigo))
> sort.dados$cod<-codigo
> rot5<-sort.dados[1,]
> for (j in 1:4){
+ i=j
+ while (i<=l){
+ area_trim<-sort.dados[which(sort.dados$ROTACAO==6410 & sort.dados$cod==i & sort.dados$trimestre==j),]
+ rot5<-rbind(rot5,area_trim)
+ i<-i+4
+ }
+ }
> limite<-length(rot5$trimestre)
> rot5_final<-rot5[2:limite,]
> l0<-length(rot5_final[1,])
> rot5_final<-rot5_final[,1:64]
> amostra_anual_2011<-rbind(dados_anual_semrot5,rot5_final)
> amostra_anual_2011[is.na(amostra_anual_2011)]<-""
```

NUTS III por ordem

```
> limite2<-length(amostra_anual_2011$NUTS.3.NG)
> NUTS.3<-0
> for (i in 1:limite2){
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==111, NUTS.3[i]<-1,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==112, NUTS.3[i]<-2,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==113, NUTS.3[i]<-3,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==114, NUTS.3[i]<-4,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==115, NUTS.3[i]<-5,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==116, NUTS.3[i]<-6,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==117, NUTS.3[i]<-7,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==118, NUTS.3[i]<-8,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==161, NUTS.3[i]<-9,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==162, NUTS.3[i]<-10,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==163, NUTS.3[i]<-11,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==164, NUTS.3[i]<-12,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==166, NUTS.3[i]<-13,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==165, NUTS.3[i]<-14,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==167, NUTS.3[i]<-15,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==168, NUTS.3[i]<-16,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==169, NUTS.3[i]<-17,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]=="16A", NUTS.3[i]<-18,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]=="16B", NUTS.3[i]<-19,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==171, NUTS.3[i]<-20,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==172, NUTS.3[i]<-21,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]=="16C", NUTS.3[i]<-22,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==185, NUTS.3[i]<-23,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==181, NUTS.3[i]<-24,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==182, NUTS.3[i]<-25,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==183, NUTS.3[i]<-26,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==184, NUTS.3[i]<-27,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==150, NUTS.3[i]<-28,
+ ifelse (amostra_anual_2011$NUTS.3.NG[i]==200, NUTS.3[i]<-29,
+ NUTS.3[i]<-30
+ ))))))))))))))))))))))))))))))}
> amostra_anual_2011$NUTS.3<-NUTS.3
```

Escalão etário

```
> esc<-0
> for (i in 1:limite2){
+ ifelse (amostra_anual_2011$IDADE_ANOS[i]>49, esc[i]<-4,
+ ifelse (amostra_anual_2011$IDADE_ANOS[i]>34, esc[i]<-3,
+ ifelse (amostra_anual_2011$IDADE_ANOS[i]>24, esc[i]<-2,
+ ifelse (amostra_anual_2011$IDADE_ANOS[i]>14, esc[i]<-1,
+ esc[i]<-0))))}
> amostra_anual_2011$esc<-esc
```

Estimativas

Número de desempregados por grupo

```
> dnut<-0
> for (j in 1:30){
+
+ desc<-0
+ for (i in 1:4){
+
+ ld<-0
+ for (k in 1:2){
+
+ cpt2<-amostra_anual_2011$CPT_PA[which(amostra_anual_2011$CPT_PA==2 & amostra_anual_2011$A4==k
& amostra_anual_2011$esc==i & amostra_anual_2011$NUTS.3==j)]
+ ld[k]<-length(cpt2)
+ }
+
+ desc<-append(desc,ld)
+
+ }
+
+ desc<-desc[2:length(desc)]
+ dnut<-append(dnut,desc)
+
+ }
> dnut<-dnut[2:length(dnut)]
> dgrupo<-matrix(dnut,30,8,byrow=TRUE)
> dgrupo<-dgrupo[1:28,]
> dgrupo_v<-c(dgrupo[,1],dgrupo[,2],dgrupo[,3],dgrupo[,4],dgrupo[,5],dgrupo[,6],dgrupo[,7],dgrupo[,8])
```

Número de inquiridos por grupo

```
> inqnut<-0
> for (j in 1:30){
+
+ inqesc<-0
+ for (i in 1:4){
+
+ linq<-0
+ for (k in 1:2){
+
+ cpt_inq<-amostra_anual_2011$CPT_PA[which(amostra_anual_2011$A4==k & amostra_anual_2011$esc==i
& amostra_anual_2011$NUTS.3==j)]
```

```
+ linq[k]<-length(cpt_inq)
+ }
+
+ inqesc<-append(inqesc,linq)
+
+ }
+
+ inqesc<-inqesc[2:length(inqesc)]
+ inqnut<-append(inqnut,inqesc)
+
+ }
> inqnut<-inqnut[2:length(inqnut)]
> inqgrupo<-matrix(inqnut,30,8,byrow=TRUE)
> inqgrupo<-inqgrupo[1:28,]
> inqgrupo_v<-c(inqgrupo[,1],inqgrupo[,2],inqgrupo[,3],inqgrupo[,4],inqgrupo[,5],inqgrupo[,6],
inqgrupo[,7],inqgrupo[,8])
```

Proporção de indivíduos registados nos Centros de Emprego - IEFP

```
> prop_iefp<-read.table("D:\\problema\\prop_iefp_anual.txt")
> prop_iefp_v<-c(prop_iefp[,1],prop_iefp[,2],prop_iefp[,3],prop_iefp[,4],prop_iefp[,5],prop_iefp[,6],
prop_iefp[,7],prop_iefp[,8])
> logit_iefp_anual<-log(prop_iefp_v/(1-prop_iefp_v))
```

Construção das variáveis dummy que entram no modelo

```
> variavel<-c(rep(1,28),rep(2,28),rep(3,28),rep(4,28),rep(5,28),rep(6,28), rep(7,28),rep(8,28))
> niveis<-1:8
> dummy<-factor(variavel,levels=niveis)

> nuts3<-rep(1:28,8)
> niveis2<-1:5
> var2<-rep(c(rep(1,8),rep(2,11),rep(3,2),2,rep(4,5),rep(5,1)),8)
> nuts2<-factor(var2,levels=niveis2)
```

Aplicação do modelo

```
> dados_anuais<-data.frame(nuts2,nuts3,dgrupo_v,inqgrupo_v,dummy,logit_iefp_anual)
> library(lme4)
> modelo8_anual<-glmer(cbind(dgrupo_v,inqgrupo_v-dgrupo_v)~dummy+logit_iefp_anual+nuts2+(1|nuts3),
data=dados_anuais,family=binomial("logit"))
> summary(modelo8_anual)

Generalized linear mixed model fit by the Laplace approximation
Formula: cbind(dgrupo_v, inqgrupo_v - dgrupo_v) ~ dummy + logit_iefp_anual +      nuts2 + (1 | nuts3)
   Data: dados_anuais
   AIC   BIC logLik deviance
 284.8 332.6 -128.4    256.8
Random effects:
 Groups Name        Variance Std.Dev.
 nuts3  (Intercept) 0.009198 0.095906
Number of obs: 224, groups: nuts3, 28
```

```
Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.995146   0.272857  -3.647 0.000265 ***
dummy2           -0.147412   0.089277  -1.651 0.098702 .
dummy3            0.097378   0.087915   1.108 0.268020
dummy4           -0.037502   0.103189  -0.363 0.716280
dummy5           -0.305939   0.083019  -3.685 0.000229 ***
dummy6           -0.405332   0.089689  -4.519 6.20e-06 ***
dummy7           -1.062018   0.077768 -13.656  < 2e-16 ***
dummy8           -1.361055   0.089835 -15.151  < 2e-16 ***
logit_iefp_anual  0.388092   0.095793   4.051 5.09e-05 ***
nuts22           -0.071225   0.084593  -0.842 0.399807
nuts23            0.219641   0.097292   2.258 0.023974 *
nuts24            0.008241   0.086730   0.095 0.924302
nuts25            0.215585   0.118122   1.825 0.067987 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) dummy2 dummy3 dummy4 dummy5 dummy6 dummy7 dummy8 lgt_f_
dummy2      -0.429
dummy3      -0.458  0.541
dummy4      -0.691  0.554  0.579
dummy5      -0.526  0.585  0.605  0.637
dummy6      -0.674  0.600  0.624  0.704  0.683
dummy7      -0.001  0.453  0.456  0.335  0.475  0.407
dummy8       0.305  0.298  0.290  0.101  0.288  0.176  0.578
logit_fp_nl  0.961 -0.296 -0.328 -0.591 -0.388 -0.553  0.169  0.465
nuts22       0.299 -0.128 -0.152 -0.263 -0.166 -0.236  0.061  0.191  0.423
nuts23       0.106 -0.056 -0.080 -0.130 -0.083 -0.115  0.032  0.098  0.209
nuts24       0.129 -0.067 -0.088 -0.147 -0.098 -0.133  0.035  0.112  0.245
nuts25      -0.097  0.011  0.000  0.004  0.003  0.010 -0.003 -0.004 -0.019
            nuts22 nuts23 nuts24
dummy2
dummy3
dummy4
dummy5
dummy6
dummy7
dummy8
logit_fp_nl
nuts22
nuts23       0.395
nuts24       0.448  0.350
nuts25       0.245  0.216  0.242
```

Totais da população por grupo sexo x escalão etário e NUTS III

```
> pop_anual<-read.table("D:\\problema\\pop_junho.txt")
> n_anual<-c(pop_anual[,1],pop_anual[,2],pop_anual[,3],pop_anual[,4],pop_anual[,5],pop_anual[,6],
pop_anual[,7],pop_anual[,8])
```

Opcional, antes da calibração: agregar por NUTS II e por NUTS III com a fórmula D=N*p

```
> ajustados<-fitted(modelo8_anual)
> total_desemp_grupo_3<-n_anual*ajustados
```

```
> total_nuts3_3<-0
> for (j in 1:28){
+ total_nuts3_3[j]<-total_desemp_grupo_3[j]
+ i=j+28
+ while (i<=224){
+ total_nuts3_3[j]<-total_nuts3_3[j]+total_desemp_grupo_3[i]
+ i<-i+28}}
> total_nuts3_3


 [1]  14595.035  25640.299  40282.081 101252.098  39137.191  18053.346
 [7]  12716.957  11764.683  23578.343  17420.104  13766.738   7029.997
[13]   1770.272  15046.365   2641.075   5240.277   3726.069   5188.316
[19]  24530.206 140473.099  59085.819  13295.536  15119.673   4851.056
[25]   7771.730   8961.801   7939.536  34332.352


> total_nuts2_3<-0
> for (i in 1:8){
+ total_nuts2_3[1]<-total_nuts2_3[1]+total_nuts3_3[i]}
> total_nuts2_3[2]<-0
> for (i in 9:19){
+ total_nuts2_3[2]<-total_nuts2_3[2]+total_nuts3_3[i]}
> total_nuts2_3[2]<-total_nuts2_3[2]+total_nuts3_3[22]
> total_nuts2_3[3]<-0
> for (i in 20:21){
+ total_nuts2_3[3]<-total_nuts2_3[3]+total_nuts3_3[i]}
> total_nuts2_3[4]<-0
> for (i in 23:27){
+ total_nuts2_3[4]<-total_nuts2_3[4]+total_nuts3_3[i]}
> total_nuts2_3[5]<-0
> total_nuts2_3[5]<-total_nuts3_3[28]
> total_nuts2_3


[1] 263441.69 133233.30 199558.92  44643.80  34332.35
```

Calibração por grupo sexo x escalão etário e NUTS II, a partir das estimativas diretas

```
> estimativas_diretas<-read.table("D:\\problema\\diretas_nuts2_grupo_anual.txt")
> estimativas_diretas<-estimativas_diretas[,1]
> estimativas_diretas


 [1] 29192.241 10187.398 19148.265  4012.524  3481.427 25939.461 10437.718
 [8] 16987.727  4380.225  2371.367 37928.527 18172.059 29343.687  6103.692
[15]  5672.052 37671.419 21661.910 28002.843  6938.927  4304.998 50903.948
[22] 18087.618 33144.786  7527.283  6803.901 47871.566 24505.801 30218.762
[29]  7775.771  4169.849 23209.925 14413.306 28891.005  5316.359  4826.550
[36] 26119.627 12084.634 19017.819  3639.990  3364.890


> pe<-0
> i=1
> while(i<=36){
+ vetor<-c(
+ rep(estimativas_diretas[i],8),
+ rep(estimativas_diretas[i+1],11),
```

```
+ rep(estimativas_diretas[i+2],2),
+ estimativas_diretas[i+1],
+ rep(estimativas_diretas[i+3],5),
+ estimativas_diretas[i+4])
+ pe<-append(pe,vetor)
+ i<-i+5}
> pe<-pe[2:225]
> pe2<-0
> j=1
> while (j<=197){
+ vetor2<-c(
+ rep(sum(total_desemp_grupo_3[j:(j+7)]),8),
+ rep(sum(total_desemp_grupo_3[(j+8):(j+18)])+total_desemp_grupo_3[j+21],11),
+ rep(sum(total_desemp_grupo_3[(j+19):(j+20)]),2),
+ sum(total_desemp_grupo_3[(j+8):(j+18)])+total_desemp_grupo_3[j+21],
+ rep(sum(total_desemp_grupo_3[(j+22):(j+26)]),5),
+ rep(total_desemp_grupo_3[j+27],1))
+ pe2<-append(pe2,vetor2)
+ j<-j+28}
> pe2<-pe2[2:225]
> lambda2<-pe/pe2
> consistentes_grupo<-total_desemp_grupo_3*lambda2
```

Agregação das estimativas por NUTS III

```
> total_nuts3_g<-0
> for (j in 1:28){
+ total_nuts3_g[j]<-consistentes_grupo[j]
+ i=j+28
+ while (i<=224){
+ total_nuts3_g[j]<-total_nuts3_g[j]+consistentes_grupo[i]
+ i<-i+28}}
> total_nuts3_g
```

```
 [1]  15407.803  27111.821  42698.237 107219.546  41476.447  19101.136
 [7]  13425.357  12396.367  22958.307  16950.959  13391.241   6836.204
[13]   1720.240  14637.067   2569.637   5104.438   3634.865   5048.732
[19]  23802.874 144090.481  60664.412  12895.881  15485.679   4963.502
[25]   7950.197   9172.343   8123.052  34995.034
```

Agregação das estimativas por NUTS II

```
> total_nuts2_g<-0
> for (i in 1:8){
+ total_nuts2_g[1]<-total_nuts2_g[1]+total_nuts3_g[i]}
> total_nuts2_g[2]<-0
> for (i in 9:19){
+ total_nuts2_g[2]<-total_nuts2_g[2]+total_nuts3_g[i]}
> total_nuts2_g[2]<-total_nuts2_g[2]+total_nuts3_g[22]
> total_nuts2_g[3]<-0
> for (i in 20:21){
+ total_nuts2_g[3]<-total_nuts2_g[3]+total_nuts3_g[i]}
> total_nuts2_g[4]<-0
> for (i in 23:27){
```

```
+ total_nuts2_g[4]<-total_nuts2_g[4]+total_nuts3_g[i]}
> total_nuts2_g[5]<-0
> total_nuts2_g[5]<-total_nuts3_g[28]
> total_nuts2_g

[1] 278836.71 129550.45 204754.89  45694.77  34995.03

> est_direta_nuts3<-read.table("D:\\problema\\diretas_nuts3_anual.txt")
> est_direta_nuts3<-est_direta_nuts3[,1]
> est_direta_nuts3

 [1]  13532.058  31518.516  42626.195 104916.419  45927.025  17465.145
 [7]  13917.125   8934.233  21922.419  16643.308  14398.989   5213.340
[13]   1440.554  17674.696   2350.374   3388.047   2549.464   3612.934
[19]  23620.031 145355.851  59399.042  16736.289  15153.768   4234.630
[25]  10312.308   7581.297   8412.770  34995.034

> est_model_based<-total_nuts3_g
```

Taxa de desemprego para 2011

```
> estimativa_empregados<-read.table("D:\\problema\\est_emprego_nuts3_anual.txt")
> estimativa_empreg<-estimativa_empregados[,1]
> estimativa_empreg

 [1] 113422.47 201914.09 244952.23 574804.31 247768.86 140723.44  96017.56
 [8]  94056.08 202283.91 157559.99 132721.89  62382.61  23000.51 148988.36
[15]  22684.96  64801.21  32716.08  43077.21 169874.77 895795.06 343396.72
[22]  96688.24 110686.82  41965.05  41621.80  76142.63  53361.37 196076.76

> taxa_desemprego<-total_nuts3_g/(total_nuts3_g+estimativa_empreg)
> taxa_desemprego

 [1] 0.11959769 0.11837884 0.14843792 0.15720791 0.14339540 0.11951313
 [7] 0.12266995 0.11644981 0.10192719 0.09713407 0.09164981 0.09876222
[13] 0.06958690 0.08945472 0.10174926 0.07301897 0.09999366 0.10490667
[19] 0.12289944 0.13856379 0.15013672 0.11768019 0.12273419 0.10576721
[25] 0.16037679 0.10751152 0.13211561 0.15144658

> taxa_desemprego<-100*taxa_desemprego
```

Gráfico da taxa de desemprego para 2011

```
xx<-1:28

jpeg(filename = "taxa_desemp_anual.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,taxa_desemprego,type="o",col="blue",xlab="nuts3",ylab="taxa de desemprego")
dev.off()
```

Taxa desemprego por nuts2

```
> est_emp_nuts2<-0
> for (i in 1:8){
+ est_emp_nuts2[1]<-est_emp_nuts2[1]+estimativa_empreg[i]}
> est_emp_nuts2[2]<-0
> for (i in 9:19){
+ est_emp_nuts2[2]<-est_emp_nuts2[2]+estimativa_empreg[i]}
> est_emp_nuts2[2]<-est_emp_nuts2[2]+estimativa_empreg[22]
> est_emp_nuts2[3]<-0
> for (i in 20:21){
+ est_emp_nuts2[3]<-est_emp_nuts2[3]+estimativa_empreg[i]}
> est_emp_nuts2[4]<-0
> for (i in 23:27){
+ est_emp_nuts2[4]<-est_emp_nuts2[4]+estimativa_empreg[i]}
> est_emp_nuts2[5]<-0
> est_emp_nuts2[5]<-estimativa_empreg[28]
> est_emp_nuts2

[1] 1713659.0 1156779.8 1239191.8  323777.7  196076.8

> tx_nuts2<-total_nuts2_g/(total_nuts2_g+est_emp_nuts2)
> tx_nuts2

[1] 0.1399434 0.1007132 0.1418023 0.1236757 0.1514466

> tx_nuts2<-tx_nuts2*100
```

Bootstrap

```
see<-0
for (i in 1:limite2){
if (amostra_anual_2011$A4[i]==1) {
ifelse (amostra_anual_2011$esc[i]==1, see[i]<-1,
ifelse (amostra_anual_2011$esc[i]==2, see[i]<-3,
ifelse (amostra_anual_2011$esc[i]==3, see[i]<-5,
ifelse (amostra_anual_2011$esc[i]==4, see[i]<-7,
see[i]<-0))))}
else
{ifelse (amostra_anual_2011$esc[i]==1, see[i]<-2,
ifelse (amostra_anual_2011$esc[i]==2, see[i]<-4,
ifelse (amostra_anual_2011$esc[i]==3, see[i]<-6,
ifelse (amostra_anual_2011$esc[i]==4, see[i]<-8,
see[i]<-0))))}}

amostra_anual_2011$see<-see

amostra_anual_2011<-amostra_anual_2011[order(amostra_anual_2011$NUTS.3),]

dados<-amostra_anual_2011

leng<-0
pe2<-0
```

```
pe<-0
total_nuts3_g2<-0
est<-0
est_nuts2<-0
taxas<-0
B<-1000

for (j in 1:B){
ordens_final<-0
pin<-1
for (i in 1:30){
leng[i]<-length(dados$CPT_PA[which(dados$NUTS.3==i)])
b<-pin:(pin+leng[i]-1)
pin<-pin+leng[i]
ordens<-sample(b,leng[i],replace=TRUE)
ordens_final<-append(ordens_final, ordens)
}
ordens_final<-ordens_final[2:length(ordens_final)]
dados_desemp<-dados[c(ordens_final),]

variavel<-0
nn<-length(dados_desemp$CPT_PA)
for (i in 1:nn) {
if (dados_desemp$CPT_PA[i]==2)
{variavel[i]<-1}
else
{variavel[i]<-0}}

dados_desemp$variavel<-variavel

tabela<-tapply(dados_desemp$CPT_PA,dados_desemp$A4,length)
tabela2<-tapply(dados_desemp$CPT_PA,dados_desemp$esc,length)
tabela33<-tapply(dados_desemp$CPT_PA,list(dados_desemp$see,dados_desemp$NUTS.3),length)
dados_f_<-replace(tabela33, is.na(tabela33), 0)
dados_f3_<-t(dados_f_)

ordem_<-c(dados_f3_[1:28,2],dados_f3_[1:28,3],dados_f3_[1:28,4],dados_f3_[1:28,5],
dados_f3_[1:28,6],dados_f3_[1:28,7],dados_f3_[1:28,8],dados_f3_[1:28,9])

dados_f4_<-as.vector(ordem_)
ninquiridos2<-dados_f4_

dados2<-dados_desemp[which(dados_desemp$CPT_PA==2),]
tabela3<-tapply(dados_desemp$variavel,list(dados_desemp$see,dados_desemp$NUTS.3),sum)
dados_f<-replace(tabela3, is.na(tabela3), 0)
dados_f3<-t(dados_f)

ordem<-c(dados_f3[1:28,2],dados_f3[1:28,3],dados_f3[1:28,4],
dados_f3[1:28,5],dados_f3[1:28,6],dados_f3[1:28,7],dados_f3[1:28,8],dados_f3[1:28,9])

dados_f4<-as.vector(ordem)

ff<-data.frame(nuts2,nuts3,dados_f4,ninquiridos2,dummy,logit_iefp_anual)

modelo<-glmer(cbind(dados_f4,ninquiridos2-dados_f4)~dummy+logit_iefp_anual+nuts2+(1|nuts3),
dados=ff,family=binomial("logit"))
```

```
ajustados<-fitted(modelo)

total_desemp_grupo<-n_anual*ajustados

pe<-0
i=1
while(i<=36){
vetor<-c(
rep(estimativas_diretas[i],8),
rep(estimativas_diretas[i+1],11),
rep(estimativas_diretas[i+2],2),
estimativas_diretas[i+1],
rep(estimativas_diretas[i+3],5),
estimativas_diretas[i+4])
pe<-append(pe,vetor)
i<-i+5}

pe<-pe[2:225]

pe2<-0
j=1
while (j<=197){
vetor2<-c(
rep(sum(total_desemp_grupo[j:(j+7)]),8),
rep(sum(total_desemp_grupo[(j+8):(j+18)])+total_desemp_grupo[j+21],11),
rep(sum(total_desemp_grupo[(j+19):(j+20)]),2),
sum(total_desemp_grupo[(j+8):(j+18)])+total_desemp_grupo[j+21],
rep(sum(total_desemp_grupo[(j+22):(j+26)]),5),
rep(total_desemp_grupo[j+27],1))
pe2<-append(pe2,vetor2)
j<-j+28}

pe2<-pe2[2:225]

lambda2<-pe/pe2

consistentes_grupo<-total_desemp_grupo*lambda2

for (j in 1:28){
total_nuts3_g2[j]<-consistentes_grupo[j]
i=j+28
while (i<=224){
total_nuts3_g2[j]<-total_nuts3_g2[j]+consistentes_grupo[i]
i<-i+28}}

est<-append(est,total_nuts3_g2)

td_nuts3<-total_nuts3_g2/(total_nuts3_g2+estimativa_empreg)

taxas<-append(taxas,td_nuts3)

total_nuts2_g2<-0
for (i in 1:8){
total_nuts2_g2[1]<-total_nuts2_g2[1]+total_nuts3_g2[i]}
```

```
total_nuts2_g2[2]<-0
for (i in 9:19){
total_nuts2_g2[2]<-total_nuts2_g2[2]+total_nuts3_g2[i]}
total_nuts2_g2[2]<-total_nuts2_g2[2]+total_nuts3_g2[22]

total_nuts2_g2[3]<-0
for (i in 20:21){
total_nuts2_g2[3]<-total_nuts2_g2[3]+total_nuts3_g2[i]}

total_nuts2_g2[4]<-0
for (i in 23:27){
total_nuts2_g2[4]<-total_nuts2_g2[4]+total_nuts3_g2[i]}

total_nuts2_g2[5]<-0
total_nuts2_g2[5]<-total_nuts3_g2[28]

est_nuts2<-append(est_nuts2,total_nuts2_g2)

}

limite<-28*B+1
est<-est[2:limite]
matriz<-matrix(est,nrow=28,ncol=B)

taxas<-taxas[2:limite]
matriz_taxa<-matrix(taxas,nrow=28,ncol=B)

limite_nuts2<-5*B+1
est_nuts2<-est_nuts2[2:limite_nuts2]
matriz_nuts2<-matrix(est_nuts2,nrow=5,ncol=B)
```

Cálculo da variância, coeficiente de variação e outros indicadores de performance

```
media_boot<-0
var_boot<-0
rb<-0
rrmse<-0
arrmse<-0
for (i in 1:28){
media_boot[i]<-mean(matriz[i,])
var_boot[i]<-sum((matriz[i,]-media_boot[i])^2)/(B-1)
rb[i]<-sum((matriz[i,]-total_nuts3_g[i])/total_nuts3_g[i])*100/B
rrmse[i]<-100*sqrt((1/B)*sum(((matriz[i,]-total_nuts3_g[i])/total_nuts3_g[i])^2))
}

arrmse<-(1/28)*sum(rrmse)
var_boot
rb
rrmse
arrmse

cv<-(sqrt(var_boot)/total_nuts3_g)*100
cv

media_boot2<-0
var_boot2<-0
```

```
for (i in 1:5){
media_boot2[i]<-mean(matriz_nuts2[i,])
var_boot2[i]<-sum(round((matriz_nuts2[i,]-media_boot2[i])^2,1))/(B-1)
}

var_boot2

cv2<-(sqrt(var_boot2)/total_nuts2_g)*100
cv2

media_boot_taxa<-0
var_boot_taxa<-0
rb_taxa<-0
rrmse_taxa<-0
arrmse_taxa<-0
for (i in 1:28){
media_boot_taxa[i]<-mean(matriz_taxa[i,])
var_boot_taxa[i]<-sum((matriz_taxa[i,]-media_boot_taxa[i])^2)/(B-1)
rb_taxa[i]<-sum((matriz_taxa[i,]-taxa_desemprego[i])/taxa_desemprego[i])*100/B
rrmse_taxa[i]<-100*sqrt((1/B)*sum(((matriz_taxa[i,]-taxa_desemprego[i])/taxa_desemprego[i])^2))
}

arrmse_taxa<-(1/28)*sum(rrmse_taxa)
var_boot_taxa
rb_taxa
rrmse_taxa
arrmse_taxa

cv_taxa<-(sqrt(var_boot_taxa)/taxa_desemprego)*100
cv_taxa
```

Alguns gráficos de indicadores de performance

```
jpeg(filename = "boxplot_cv_anual.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
boxplot(cv)
dev.off()

se_boot<-sqrt(var_boot)
se_boot

xx<-1:28
jpeg(filename = "se_anual.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,se_boot,type="o",,col="blue",xlab="nuts3",ylab="desvio padrão est. model based")
dev.off()

jpeg(filename = "cv_anual.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,cv,type="o",col="blue",xlab="nuts3",ylab="cv est. model based")
dev.off()
```

```
#load(workspace cvs anuais direto)
cv_direto<-cv_pais_2
cv_pais_2

se_direto<-(cv_direto/100)*est_direta_nuts3

jpeg(filename = "se_direct_model_anual.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(se_direto,se_boot,xlab="desvio padrão est. direta", ylab="desvio padrão est. model based",col="blue")
abline(a=0,b=1,col="red")
legend(5000,3700, pt.bg="white", lty=1,"reta y=x",col="red")
dev.off()

jpeg(filename = "cv_direct_model_anual.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,cv_direto,type="o",col="blue",xlab="Nuts3",ylab="CV")
lines(cv,type="o",pch=21, lty=1,col="red")
legend(18,52,pch=21, pt.bg="white", lty=1,c("est. direta","est. model based"),col=c("blue","red"))
dev.off()

jpeg(filename = "rrmse_anual.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
plot(xx,rrmse,type="o",,col="blue",xlab="nuts3",ylab="rrmse est. model based")
dev.off()

jpeg(filename = "boxplot_rb_anual.jpg", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA)
boxplot(rb)
dev.off()
```